MOLECULAR MECHANICS
FORCE FIELD OPTIMISATION USING
PARALLEL COMPUTERS

A Thesis presented for

the degree of Doctor of Philosophy

in the

Faculty of Science

of the

University of Glasgow

by

Paul Rex Edgington

September 1992

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Ever since the discovery that the bond lengths and angles between the same atomic types vary little between different molecules chemists have been using models of molecules to aid them in their research. This technique is known as molecular modelling and is now usually performed with the aid of computers. One of the main areas of interest in molecular modelling is that of molecular mechanics calculations.

Molecular mechanics (MM) is the use of empirical equations and their associated parameter to determine the energy of any particular molecular conformation and the use of that energy in helping to predict such things as the minimum energy conformation. The equations and the parameters are together described as the MM force field. One drawback of molecular mechanics is that in certain situations where its application would be useful there is an absence of reliable parameters so giving rise to the situation where calculations are often performed using 'guesstimates' for parameter values. To improve this situation an investigation into the possible forms of a general purpose compact force field was undertaken.

There exist several force fields that reduce the number of parameters required by simplification of the situations where a parameter may be applied, for example a torsional barrier would be defined in terms of the central two atom types only rather than the more rigorous case where the atom types of all four atoms in the torsion angle would be used to decide which parameter is to be used. Some work had also already been done on reducing the number of parameters by calculating some by using other empirical formulae, an example of this being the calculation of bond stretch constants from their associated bond length parameters. These equations usually require some optimisable 'constants' but, in general, the technique resulted in an overall reduction in the number of parameters that are required to be optimised.

What is not obvious is which of these methods are best at reducing the number of parameters without having an excessive effect on the accuracy of the results produced by the final force field. For each of these published force fields it is possible to adapt an energy minimisation program to calculate the minimum energy structures of known molecules and then compare the calculated results with experimental data. This will give us an overview of the accuracy for that force field but will not show which of the methods used within that force field have been the most effective. As MM is an empirical approach there are force fields which have basically the same form but have different parameters and it is usually unwise to interchange the parameters between them without reoptimising the whole force field.

At the start of the project the methods used to optimise force fields were slow and the production of even an optimised force field for alkanes could take many years

even after the form of the equations had been decided. It was obvious that using normal methods of optimisation it would take far too long to implement a system to vary the methods of parameter reduction and then reoptimise the force field in each case to see its effects. To overcome this problem it was decided to produce the required force fields by computer optimisation. This was made possible in mainly by the recently available parallel computing power of the Inmos transputer chip.

The method chosen the controlling computer program to alter the parameters and so could be left for long periods to produce an optimised force field in a fraction of the time that would previously have been required.

The first studies were done on small sets of alkene data as initially the processing power was limited. These initial studies culminated in investigations using a set of 50 alkene structures. These showed that a highly reduced force field is a viable option for alkenes, however alkenes are not very representative of all the atom types that will be needed so it was decided to introduce some more atom types before deciding on the final form of the force field. To this end a set of 109 structures was constructed which contained the following atom types: H, C_{sp2}, C_{sp3}, O_{sp2}, Cl, Br and F.

After considering the results of optimisations with this set of structures the form of a highly reduced parameter force field was decided upon. This force field was used in a limited study using a set of 243 structures with 25 different atom types.

The following table summarises the results of this optimisation. Conformer refers to energy difference measures between conformers, 'Average Diff' is the average difference between the experimental value and that predicted by the force field and 'Average Error' is the average experimental error for that property type.

Property Type	No. of Properties	Average Diff	Average Error
Distance	2934	0.0581Å	0.0380Å
Valence Angle	2910	2.4087°	2.0963°
Torsion Angle	3328	5.3090°	5.5671°
Conformer	17	0.5292 Kcal Mol^{-1}	0.3084 Kcal Mol^{-1}
Heat of Formation	46	1.2911 Kcal Mol^{-1}	0.2499 Kcal Mol^{-1}

Although not totally conclusive this study indicated that a highly reduced force field would be a valuable addition to the range of force fields available to the molecular modeller, as while being not as accurate a fuller force field it would have the significant advantage of covering, with reasonable accuracy, those situations that are not parameterised in the more specific force fields.

GREER'S THIRD LAW:

A computer program does what you tell it to do, not what you want it to do.

ACKNOWLEDGEMENTS

I wish to acknowledge all the help given to me by Dr. David White at Glasgow and Dr. Mike Lees at Ciba Geigy in Manchester. I would also like to thank everyone else who helped through my period of research including Mr. Noel Ruddock for his help with occam, Dr. Guy Grant and Dr Paul Yates for their expert advice on molecular modelling and Dr. Kelvin Tyler for his all his assistance.

I would also like to thank Dr. Olga Kennard and Dr. Frank Allen from the Cambridge Crystallographic Data Centre for their support.

Financial support from Ciba Geigy Industrial Chemicals of Manchester is gratefully acknowledged.

Finally, I would like to thank my parents who had to put up with me not having a real job for so long.

Cambridge, 1992 P.R.Edgington

Table of Contents

Figures and Tables

Figures

Chapter 7

Tables

Chapter 8

CHAPTER 1

Molecular Modelling

1.1 Introduction

The term molecular modelling covers an extensive area of topics ranging from simple ball and stick representations of small molecules to complicated molecular dynamic's simulation studies of many molecule systems. The common factor is that the molecule or molecules under investigation are considered as a set of co-ordinates to which the appropriate equations are applied and that the relevance of any results is dependent on how accurately the equations used represent (i.e. model) the appropriate effects, in the real world, on the real system.

Although physical models are still used at times, the vast majority of molecular modelling done today is performed using computers. These modelling programs can be divided into two groups.

The first of these are the single dedicated programs[1]. In these cases the program will usually start by loading a file containing the structural data on a molecule. The program will then proceed to perform a single function such as energy minimisation or calculation of partial charges before terminating having recorded any required results.

The other approach is to have a package that integrates many different options[2]. This means that it is possible to stay within the single package while performing a wide range of operations. With the increase in computer power and memory size the integrated packages are becoming more useful and thus more used.

1.2 COMMET

An example of an integrated package is the COMMET system (**CO**ncurrent **M**olecular **M**odelling **E**nvironment on **T**ransputers) which has been developed within our laboratory. This package was developed from the VAX program COGS[3].

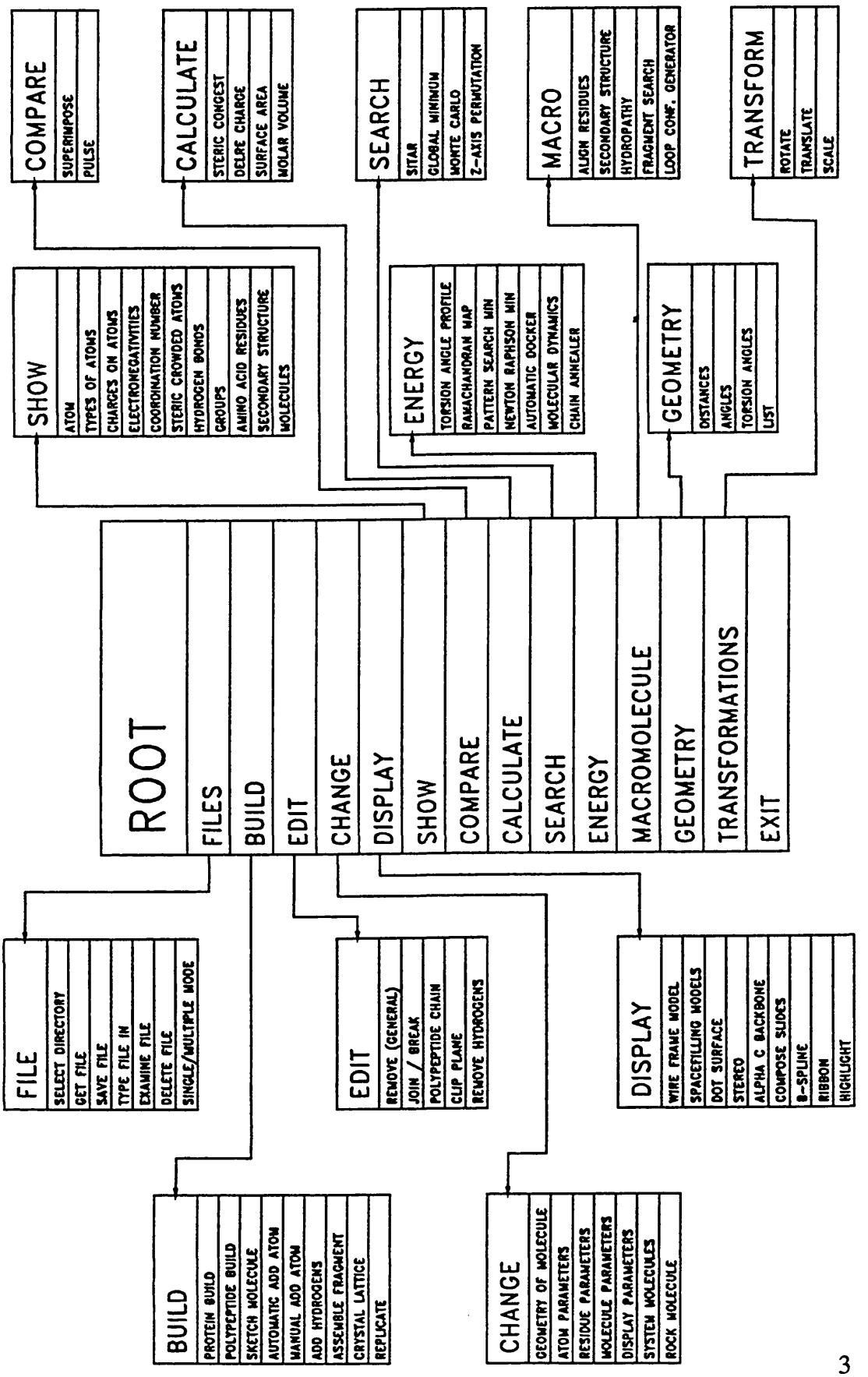
COMMET is a graphics based, menu driven package and Figure 1.1 shows a diagrammatic layout of the menus available in the COMMET system and gives an idea of the range of options that can be performed from within the system.

It is useful to examine the various functions within such a package to obtain an overview of the options covered by the general heading of 'Molecular Modelling'.

1.2.1 Files

This menu contains all the file operations available from within the system. These include reading, writing, inspecting and erasing of files, changing the current directory and the ability to decide if any new structure being loaded will replace the currently resident structure or be loaded alongside the current structure.

Figure 1.1 COMMET Modelling Package



1.2.2 Build

This menu contains the options used in the construction of molecules. Building can occur in two ways, either by adding atoms or by the addition of amino acids to build up polypeptide chains. There is also a **sketch molecule** option where the user can sketch the required molecule on the screen with a mouse and indicate which atoms are above and below the plane. This option will then invoke an energy minimiser from which the final structure can be obtained. The **protein build** option allows the user to edit the amino acid sequence and the three dimensional structure of a protein. For smaller molecules the option **assemble fragment** allows the user to join two molecules by the elimination of two hydrogen's.

1.2.3 Edit

The edit menu is mainly used for the removal of all or part of molecules, and for joining or breaking bonds.

1.2.4 Change

There are two parts to this menu, the first involves changing certain properties relating to molecule itself, for example, such things as bond length, bond angle, atom types and atom charges. The other menu items are for making changes to the default program parameters, for example, the colour of the backbone, whether the depth cue option is on or off and the hydrogen visibility.

1.2.5 Display

This menu allows the user to change the way that the molecule is displayed on the screen. Initially the molecule will be displayed as a simple wire framed model. The following other options are available:

Advanced wire frame (displays bond multiplicity).

Simple space filled (simple circles).

CPK surface (Intersecting spheres of van der Waals radii).

Dot Surface (as CPK but with dots rather than full spheres).

Stereo (red and green images displayed to produce a stereo image when viewed with the appropriate coloured glasses).

Alpha carbon backbone (single line going along alpha carbons in a protein or polypeptide).

Beta-spline (as for the alpha carbon backbone but using a beta-spline smoothing technique to produce a smooth curve).

Ribbon (as for beta-spline but with several splines side by side to produce a ribbon effect).

It is also possible in this menu to produce screen displays which are suitable for being photographed and to highlight selected areas of the molecule.

1.2.6 Show

In the show menu the user can highlight various selected groups, display several different atomic properties and determine which atoms are likely to be sterically crowded or involved in hydrogen bonding.

1.2.7 Compare

This menu allows the superimposition of molecules and then, if required, to pulse between them to help show where any differences and similarities are.

1.2.8 Calculate

This menu contains the options for calculating the **steric congest**[4] at any atom or pairs of atoms where it is possible for a stereoscopic reaction to take place. **Delre charge**[5] will calculate the charge on a molecule using the Delre method. The menu also contains options to calculate both the surface area and volume of any selected molecule.

1.2.9 Search

This menu contains routines that attempt to find the global minimum energy position of molecules using different methods. **SITAR**[3] is used to find the global minimum of side chains in polypeptides. The **Global minimum**[3] option is used for cyclic polypeptides. The **Monte Carlo**[6] routine will attempt to use the Monte Carlo technique and can be applied to any molecular situation. The **Z-axis permutation** is used for small molecules that contain one or more rings.

1.2.10 Energy

This menu contains most of the other options that are needed to calculate the energy of a system. The options are as follows:

Torsion angle profile and **Rammachandaran map**[7] calculate energy profile plots for rotations about one and two torsion angles respectively.

Pattern search[8] and **Newton-Raphson**[8] minimisation's are energy minimisation routines that use the specified techniques for minimisation (see chapter 2).

Auto docker attempts to dock a molecule into a specific site on another molecule. **Molecular dynamics**[8] is the simulation of the molecule above absolute zero so that atoms will have vibrational motion. **Chain annealer** is used when an alteration has been made to a protein chain and involves minimisation only along the section of the

chain where the alteration has taken place.

1.2.11 Macromolecules

The options in this menu are related to the order of the residues in a protein. **Align residues** calculates the homology between two different proteins. **Secondary structure** analyses a chain of residues and plots the residues probable propensity towards forming alpha helixes or beta sheets. **Hydropathy** shows how the hydrophobic and hydrophilic character of the chain varies along its length. **Fragment search** and **Loop conformation generator** are used to generate the structure of a part of the chain that has been altered. **Fragment search** looks through a database of structures trying to find a similar loop in a comparable situation. The **Loop conformation generator** examines all the possible loop conformations, using filters such as distance tolerance and torsion angles to reduce the number of selected loops.

1.2.12 Geometry

The options in this menu allow the user to measure simple geometric properties such as bond length, valance bond angle and torsion angle.

1.2.13 Transformation

This menu allows the user to manipulate the position of any molecule by translation and rotation and to scale the size of the molecules on the screen.

1.3 Summary

Within such a modelling package as COMMET there is a wide range of operations that can be applied to a range of different molecular situations. It is advantageous at this stage to try to make a degree of distinction between different sorts of operations.

One way of partitioning the operations is to look at the degree of extra parameterisation that is required for an option to function properly. For the majority of the operations there is either no requirement for extra parameters or the number of extra parameters needed will be small and relatively easy to obtain. Examples of this sort of operation are the molecular transformations such as rotations that require no extra information and the dot surface and CPK space filled displays where only the van der Waals radii of the atoms' types to be displayed need to be known.

Above these options are those which require a significant degree of parameterisation and in many of these cases the production of the required parameters is not a trivial operation. The prime examples of these types of operations are the

charge calculation option and the diverse options such as energy minimisation and molecular dynamics that require the utilisation of a method of empirical energy calculation. For the implementation of these options there are two tasks that need to be completed. Firstly there is the production of the actual program code, but as well as this there is the need to either find or produce the relevant parameters so that the option can be used.

Thus covering the whole sphere of molecular modelling it can be seen that there are three area's where a significant undertaking of work is required before meaningful results can be produced.

1) The production of the programs.

2) The utilisation of the programs.

3) The production of the parameters to be used within the programs.

It is this third part that will be the main subject of the work described here.

1.4 References

1 (a) **MM2**, N.L.Allinger and Y.H.Yut, *Molecular Mechanics*, 1980,QCPE Program No. 395.

(b) **CHARGE2**, R.J.Abraham, G.H.Grant, B.Hudson and P.E.Smith, University of Liverpool, Liverpool, UK.

2 (a) **BIOGRAF**, BioDesign inc., Pasadena, CA, USA.

(b) **QUANTA**, Polygen Corporation, Waltham, MA, USA.

(c) **Chem-X**, Chemical Design Ltd, Oxford, UK.

(d) **ALCHEMY**, Tripos Associates Inc., St Louis ,MO, USA.

3 D.N.J.White and J.E.Pearson, *J. Mol. Graph.*, 1986, **4**, 134.

4 K.T.Wipke and P.Gund, *J. Am. Chem. Soc.*, 1976, **98**, 8107.

5 G.Del Re, *Biochem. Biophys. Acta*, 1963, **75**, 153.

6 N.Go and H.A.Sheraga, *Macromolecules*, 1978, **11**, 552.

7 G.N.Ramachandran and V.Sasisekharan, *Adv. Protein Chem.*, 1968, **23**, 283.

8 D.N.J.White, *Computer aided Molecular design*, Oyez, London, 1984, 73.

CHAPTER 2

Molecular Mechanics

2.1 Introduction

Molecular mechanics can be considered as a technique for calculating many properties of molecular systems, based on the ability to determine an estimation of the energy for such systems in any atomic configuration. When calculating the energy it is assumed that all the interactions within a molecule can be treated in an empirical manner.

Even though molecular mechanics is empirical in nature it can be justified, to some extent, from quantum mechanics. This is done by examining the Born-Oppenheimer approximation. This states that in quantum mechanics it is possible to separate the motion of the nuclei in a molecule from the motion of their associated electrons, with little effect on the calculated results. This is used to find the electronic structure of a molecule by considering the nuclei to be fixed in a given configuration. It is equally valid though, to investigate the motion of the nuclei. Here it can be considered that the surrounding electron density leads to the various interactions represented in molecular mechanics by the potential functions.

The other main supposition in molecular modelling is that the interactions between atoms within a molecule can be divided into various distinct types. Examples of the types of interaction are bond length stretching, valence bond angle bending, torsion angle twisting and non-bonded repulsion or attraction. A separate empirical equation is used to calculate the energy for each interaction type.

The terms used for each of these interaction types and the empirical parameters that these terms need are what is known collectively as the force field.

The total structural energy of the molecule will be a simple sum of the energy calculated for each of the terms. This is known as the steric or strain energy. Thus the total steric energy E_s of a molecule can be given by:

$$E_s = E_b + E_a + E_t + E_{nb} + E_c \; (+ E_x)$$

Where E_b is the energy relating to bond stretching or compression, E_a is the energy of valence angle distortion, E_t is the component relating to the energy from torsional barriers, E_{nb} is the non-bonded contribution from the van der Waals potential and E_c is the coulombic energy arising from charge interactions. E_x represents the possibility that other terms may be needed such as the out of plane bending of planar atom types and possible cross terms that could be added to increase accuracy.

For each term in the force field there will usually be several possible equations that could be used and in each of these cases there will usually be several different ways that the required parameters may be selected.

For the majority of these terms the parameters consist of a "natural" or "strain

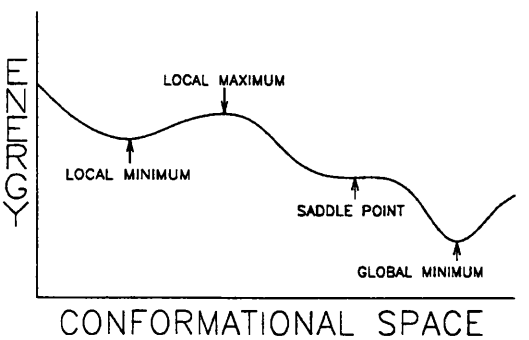
free" value and one or more force constants that determine how difficult it is to deform the property from this "strain free" value. The interaction types not using this approach are the torsional twist where periodicities and barrier heights are used and the non-bonded interaction which uses separate terms to cover attraction and repulsion plus a single term for coulombic interaction.

The basic idea of using an empirical force field had been proposed as early as the 1930's[1] but serious attempts to use molecular mechanics where not made until 1946[2-4]. Due to the lack of reliable information on which to base the parameters and the difficulties involved in doing large calculations at the time, the more widespread use of molecular mechanics did not start until the 1950's[5]. Since then its importance has increased steadily with the growth in computer power and with the increase in the number of computational approaches that have been devised and implemented.

For a molecule with N atoms it is possible to imagine a 3N dimensional surface that describes the energy of the molecule in its electronic ground state as it is affected by the values of the 3N co-ordinates. This surface is usually called the Born-Oppenheimer surface or the potential energy hypersurface. Depending on the symmetry of the molecule under investigation there will be one or more locations on this surface that will correspond to the lowest energy position of the molecule. Thus, as well as a method to determine the energy at any location there is also a need to manipulate the molecular co-ordinates to allow for a search of the Born-Oppenheimer surface so that the position or positions of minimum energy can be found.

This is done using a minimisation technique that will usually examine the forces on the atoms and then apply transformations to the co-ordinates in an attempt to obtain a configuration with smaller average forces. This process will be repeated iteratively until a stable structure is reached. One of the main drawbacks of these optimisation methods is that the minimum is found by heading towards the nearest local energy minimum and that this is generally not the global minimum for the given structure. As well as this, some minimisation techniques can terminate at an energy maximum or get caught at a saddle point in the potential energy surface (figure 2.1).

Figure 2.1 Likely Elements on a Potential Energy Surface



Some of the various methods of energy minimisation are described later. In most cases the Newton-Raphson technique is used as it provides a rapid method of reaching the local minimum. It is not, though, the optimum technique when the initial structure is highly distorted from the minimum energy position. Because of this some molecular mechanics programs initially use a different technique such as steepest descent to get closer to the minimum before using the Newton-Raphson method.

One of the problems with the Newton-Raphson technique arises with large deformations and in some cases it can result in a situation of increasing oscillation for the affected atomic positions. To counteract this a variation on the line search technique has been implemented which recognises the start of such oscillation and counteracts it by reducing the maximum atomic shift until a reduction rather than an increase in the energy is obtained.

2.2 Why Use Molecular Mechanics?

With all the problems involved with the production and use of a molecular mechanics force field there is a need to consider why molecular mechanics should be attempted. Results can, after all, be obtained both from various experimental techniques and from *ab initio* quantum mechanics calculations which does not require any empirical parameters.

What then are the advantages of molecular mechanics that have resulted in its rising popularity over recent years?

The main reason is the speed and convenience with which it is possible to produce the required answers. This is especially true because of the large range of data that can be obtained from molecular mechanics calculations. For example, if an investigation of a new molecule is required then, to get the structure alone experimentally, would require that the molecule be synthesised, followed by the need to determine the molecular conformation using one of several methods, such as x-ray diffraction.

Thus each result obtained from molecular mechanics calculations, such as an estimate for the heat of formation, would, if determined experimentally, require its own experimental procedure. So in many cases such calculations can save a considerable amount of time, effort and money. This is, however, not to say that such experimentation should cease. The whole basis of molecular mechanics is that it is an empirical technique and so will never totally replace the more accurate experimental approaches. Indeed it is the expanding database of such experimental data that allows for the continually improving quality of molecular mechanics force fields.

What then of quantum mechanical calculations? They too can be used to find the minimum energy of a molecule by computation alone. The principal problem here is that, when doing a quantum mechanics calculation there is always a trade off of accuracy against calculation time depending on which basis set of atomic wave functions is chosen. With the simpler wave functions the calculations are done more rapidly, if still orders of magnitude slower than molecular mechanics, but a simple basis set is itself an approximation and so the results themselves are also approximate[6]. The use of a more accurate basis set will improve the accuracy of the results but this will also substantially increase the computational time required. Another advantage with molecular mechanics is that the computation time increases roughly with the square of the number of atoms involved, whereas, with quantum mechanics, it increases with the fourth power of the number of atoms. This means that, though molecular mechanics calculations are regularly performed on systems of hundreds of thousands of atoms, such systems are still out of the applicable range of *ab initio* quantum mechanics calculations, even on the largest supercomputers.

There are also several popular quantum mechanics techniques that neglect specific orbital overlaps to speed up the calculation but these bring with them the need to use some empirical parameters in an attempt to make up for the resulting deficiencies, hence the term semi-empirical that is given to these methods[6].

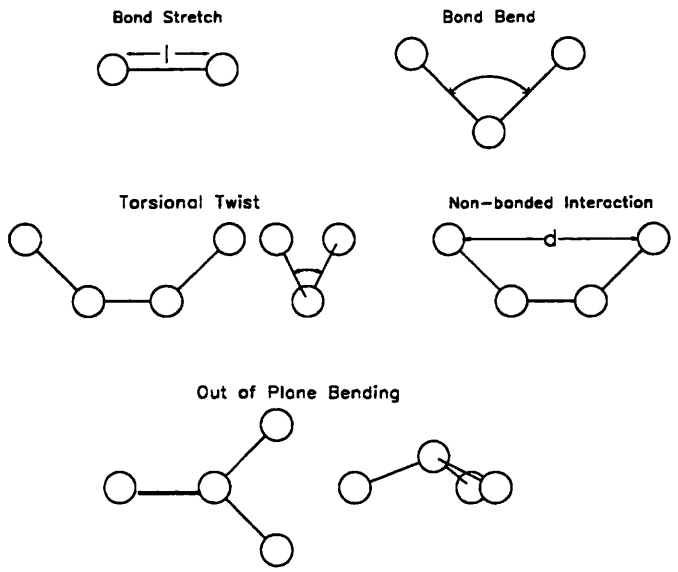
Even with the resultant increase in speed none of these methods come close to the accuracy of molecular mechanics within the same computational time scale.

Quantum mechanics calculations are still very useful, especially when investigating situations that are not reliably parameterised for molecular mechanics or when studying reactions, which are difficult or impossible to simulate using molecular mechanics.

2.3 Force Fields

As previously stated the force field consists both of the form of the potential functions used to calculate each energy component and the related parameters. Figure 2.2 shows the main interaction types for which potential energy functions will be required.

Figure 2.2 Interaction Types



Looking at each interaction type in turn:

2.3.1 Bond Stretching

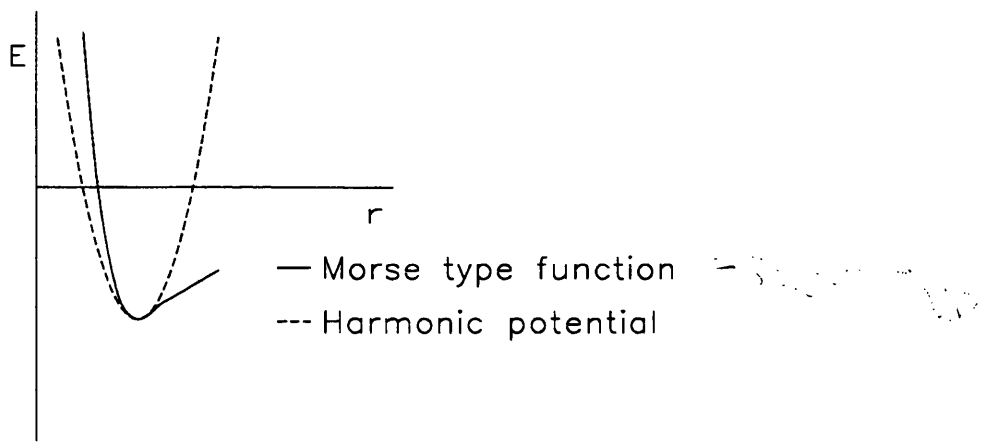
The potential function most commonly used for the calculation of lengthways deformation of bonds, E_b is:

$$E_b = \sum_i \tfrac{1}{2} K_i (l - l_0)^2$$

Where K_l is the force constant, l is the current bond length, l_O the strain free bond length and the sum is over all the bonds in a molecule.

A simple equation such as this can be used because to attempt to stretch or compress a bond is energetically expensive and so the value of l is unlikely to be far from l_O. So, though the actual bond energy varies as a Morse curve[7], it can, for a limited range, be approximated to a simple harmonic Hookean function (figure 2.3).

Figure 2.3 Morse and Harmonic Curves for Bond Lengths



A more complicated form of this equation is used in the recently published MM3[8] alkane paper which uses the following form for each bond length:

$$71.94K_s(l-l_0)^2[1-2.55(l-l_0)+(7/12)2.55(l-l_0)^2]$$

These extra terms allow for the few situations where the bond length is greatly distorted. This use of extra terms is shown in the next section on the valence angle bending interaction.

2.3.2 Valance Angle Bending

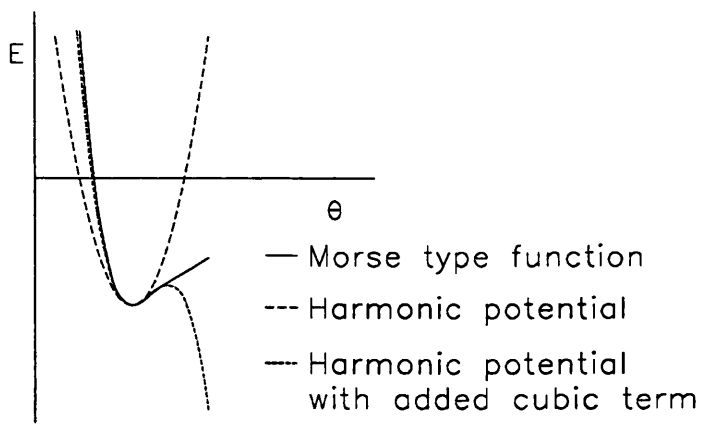
In the case of valance angle bending the force constants are smaller giving rise to shallower energy curves and so allowing greater distortion away from the strain free value. This means that the simple harmonic equation, as shown below, is less realistic than in the case of the bond stretching.

$$E_a = \sum_{\theta} \tfrac{1}{2} F_\theta (\theta - \theta_0)^2$$

This situation can be improved by adding a single anharmonic correction term[9] (figure 2.4) to give the following equation.

$$E_\theta = \sum_{\theta} \tfrac{1}{2}\left(F(\theta-\theta_0)^2 - F_\theta^1(\theta-\theta_0)^3\right)$$

Figure 2.4 Morse, Harmonic and Harmonic + Cubic Curves



This works well, except in the few cases where the angle starts off being greatly deformed from the strain free value. A situation where this could easily occur is during the energy minimisation of a molecule that has been sketched into a molecular modelling package. In these cases the cubic correction term can become dominant over the squared term. Because of this the total energy will decrease if the angle is further deformed away from the strain free value, an obviously unrealistic simulation. To prevent this from occurring an extra term can be added, which will, in

these extreme cases have a greater effect than the cubic term and so force the angle back towards a more reasonable value. The fifth power is often chosen because it does not drastically increase the calculation required as it can be produced from the product of the square and the cubic terms.

$$E = \sum_{\theta} \tfrac{1}{2}F(\theta - \theta_0)^2 - F_\theta^1(\theta - \theta_0)^2 + F_\theta^2(\theta - \theta_0)^3$$

Where F_θ, F_θ^1 and F_θ^2 are force constants, θ is the bond angle and θ_O is the appropriate strain free bond angle for the situation under consideration. As the F_θ^2 term is only present to correct for large deformations this constant is usually fixed at some small fraction of the F_θ^1 term.

This equation can be considered as being produced by taking selected terms out of a Taylor Expansion, in MM3[8] the Taylor Expansion is taken fully up to the 6th term and the parameters are all given as a proportion of the first angle bending constant:

$$E_a = 0.021914(k_\theta)(\theta - \theta_0)^2[1 - 0.014(\theta - \theta_0) + 5.5\times10^{-5}\times(\theta - \theta_0)^2$$
$$-7.0\times10^{-7}(\theta - \theta_0)^3 + 9.0\times10^{-10}(\theta - \theta_0)^4]$$

2.3.3 Torsional Strain

The next term to be considered is that relating to the torsional energy E_t. It had long been thought that the need for a torsional energy term in a force field was due only to the deficiencies in the other parameters, most specifically the non-bonded interactions, and that if these could be optimised properly then this term would not be needed. There have been, however, several attempts to devise force fields that do not contain a torsional term but these have failed even in the simplest cases as they could not, for example, even reproduce the correct internal rotational barrier height for ethane[10] without giving incorrect answers for other properties such as angles.

The need for the torsional term is usually rationalised as being required to model the interactions between bonds that are not adequately covered by the interactions between the atoms.

Initially the torsional energy was calculated using a term simply relating to the periodicity of the central bond so, for ethane, the equation would be:

$$E_t = V_3(1 - \cos(2\omega))$$

Where V_3 is the barrier height and ω is the measured torsion angle. However, for certain torsion angles this is not sufficient and it is has been found[11] that for certain situations an additional one fold term is required, giving a final equation of:

$$E_t = \sum_{\omega} \tfrac{1}{2}[V_n(1 + s\cos(n\omega)) + V_1(1 + s\cos(\omega))]$$

Where the summation is over all the torsion angles around all the bonds. V_n and V_1 are the n-fold and one-fold components of the barrier to free rotation, ω is the torsion angle, n its periodicity, and s = +1 for a staggered minimum energy configuration (e.g. ethane) and s = -1 for situations where the minimum energy configuration is eclipsed (e.g. ethylene). In most cases the V_1 parameter is set to zero but has been found to give a better fit to experimental data when it is used for the C_{sp3}-C_{sp3}-C_{sp3}-C_{sp3} torsion angle and the O_{sp2}=C_{sp2}-N_{amide}-H torsion angle (e.g. in an amide link).

A slightly different system has been used in Allingers MM2[12] and MM3[8] where three contributions to the overall barrier are given, relating to one fold, two fold and three fold periodicities.

2.3.4 Non-Bonded Interactions

There are two types of non-bonded interactions, the Coulombic interactions between the charges on the atoms and the van der Waals interactions between the atoms themselves.

2.3.4.1 Charge Interactions

E_c is the energy component relating to the charge interactions.

A real molecule consists of the positively charged atomic nuclei surrounded by the negatively charged electrons in their appropriate orbitals. This results in a charge distribution that extends throughout the volume of the molecule. The calculation of such a distribution and from that the calculation of the total Coulombic energy is very computationally intensive and as such is not particularly feasible for the time scale of molecular mechanics calculations.

Instead force fields containing a charge term use either a system where the charge is distributed as point charges on the atom nuclei or one that assign dipoles along each bond.

With the dipole approach the usual method is for each pair of bonded atoms to be given a dipole depending on the atoms in the bond. This means that their values are easy to assign but require all the relevant values for the dipoles to be known. The dipole values also do not take into account the effect that any other atoms bonded to either of the relevant atoms might have.

The calculation of point charges will usually take these effects into account but the calculation of a charge distribution is itself not a simple problem and can require a substantial amount of computation. This point charge distribution, though,

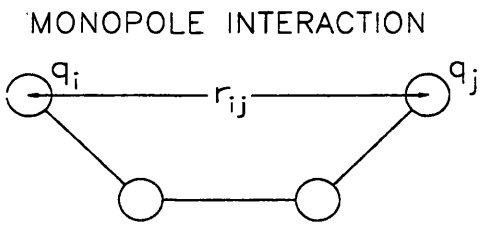
needs only be calculated once as the charges are usually calculated just from the connectivities and the atom types and thus should be equally valid whatever the actual atomic positions.

In the point charge case the energy can be obtained from the sum of the pairwise interactions between all the possible combinations of monopoles, i.e.:

$$E_c = 332 \sum_r \frac{q_i q_j}{D r_{ij}}$$

Where q_i and q_j are the charges (in units of electronic charge), on the atoms i and j separated by the distance r_{ij} and D is the dielectric constant. The scaling factor of 332 converts the energy to units of Kcal per mole. Figure 2.5 shows a single pairwise monopole interaction.

Figure 2.5 Single Pairwise Interaction

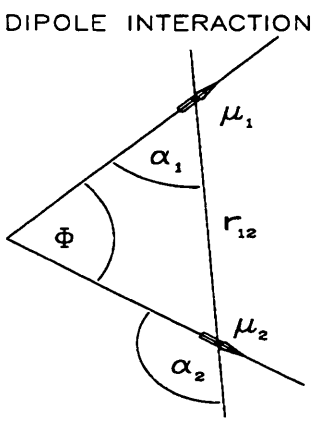


In the dipole case the total energy for the charge interaction is obtained from the sum of the interactions of the pairs of dipoles[13], i.e.:

$$E_c = \sum_r \frac{\mu_i \mu_j}{D r_{ij}} \left(\cos\Phi - 3\cos\alpha_i \cos\alpha_j \right)$$

Where D is the dielectric constant, r_{ij} is the separation of the dipoles, Φ is the angle between the dipoles, μ_i and μ_j are the values of the dipoles and α_i and α_j are the angles each dipole makes to a line connecting them. Figure 2.6 shows a single dipole interaction.

Figure 2.6 Single Dipole Interaction

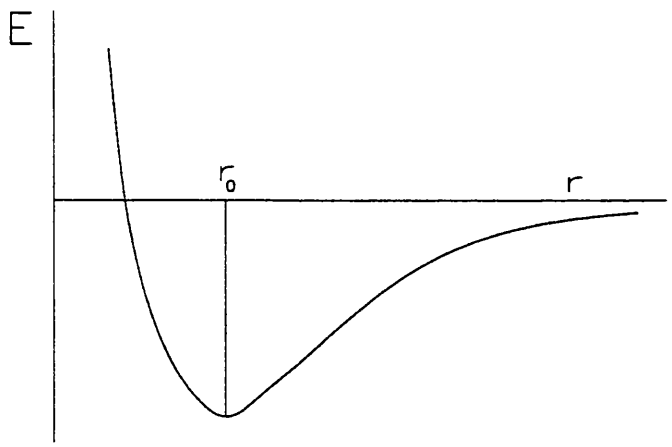


The value used for the dielectric constant is a continuing point of dispute. Some force fields[12] use a fixed dielectric constant, usually 1, which is claimed to accurately represent the interaction of the collection of point charges being considered. Other force fields[22] have been optimised where it is considered that giving the dielectric constant a fixed value, such as 1, will tend to overestimate the energy of the charge interactions. In these force fields the dielectric constant is linked to the distance of seperation r. When this is done an extra scaling factor may also be used so that, at the closest likely distance for the nuclei, the effective dielectric constant will be close to 1. When considering which to use it has to be remembered that as molecular modelling is an empirical technique each approach is equally valid. In the studies described here in the method of linking the dielectric constant to the distance has been used.

2.3.4.2 Van der Waals Interactions

Whereas with all the other terms the choice of potential function is quite often just between the level of complexity of a particular form, the van der Waals interaction is different in that there are a number of alternative forms for the equations. All of these equations attempt to reproduce the typical shape of the van der Waals interaction (figure 2.7).

Figure 2.7 Typical van der Waals Interaction



The following are a selection of the possible options.

Lennard-Jones Potential[14]

$$E_{nb} = \sum_{r} \left[Ar^{-n} - Br^{-6} \right]$$ Where n is usually 12,10 or 9.

Where A and B are the van der Waals interaction constants and their values depend on the atom types involved in the interaction. The greater the value of the power n used, the harder (i.e. steeper) will be the short range repulsion.

Buckingham (6 - exp) Potential[15]

$$E_{nb} = \sum_r F_r \left\{ \frac{-2}{\alpha^6} + \exp[12(1-\alpha)] \right\}$$

Where a = r/(r_1^* + r_2^*) and r_1^*,r_2^* are van der Waals interaction constants and relate to the van der Waals radii. The exponential relates to the repulsive part of the interaction and the factor 12 gives a measure of the steepness of this repulsion and so could also be considered to be a optimisable parameter.

The use of a simple pairwise potential has been criticised[16] because it neglects many body effects and that the van der Waals interactions would be affected by the electron density of any other atoms between the two atoms in question, in a similar way to the coulombic interaction. Despite this it is still used as these effects appear to be insignificant compared to the advantages obtained in calculation time.

A greater problem with the van der Waals interaction is that in both of the above equations the atoms are assumed to be spherical. This generalisation results in two main problems.

The first is that some types of atoms, such as oxygen, have lone pairs of electrons with the result that, in the real molecule, the repulsion between such a lone pair bearing atom type and another atom will vary depending on the location of the lone pair electrons with respect to the atom centres. That is, the repulsion will be greater when a lone pair is between the atoms' centres than when they are not. It is possible to simulate this situation by adding lone pairs as pseudo atoms but this will necessarily result in an increase in the number of parameters required in the force field.

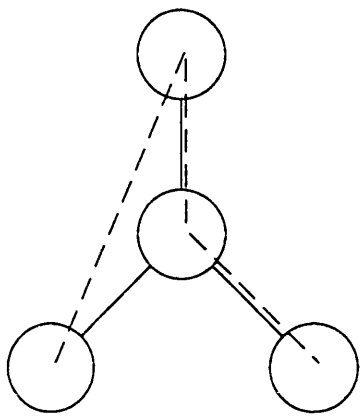
A second problem is that bonded hydrogen is found to have an electron density centred not at the hydrogen nucleus but instead it shifted along the bond towards the other atom in the bond. Some force fields attempt reproduce this effect by moving the centre of the hydrogen atoms about 10% along the bond when calculating the non-bonded interactions[17,18]. This has been found in these cases to reproduce the crystal packing of hydrocarbons better than when the centres are taken at their normal positions. This approach will increase the computation time as the "new" atomic centre for each hydrogen will need to be re-calculated at every point required.

2.3.5 Out of Plane bending

Many force fields also include a separate term to account for the energy increase with out of plane bending (pyramidalization) of a trigonal planar system

(e.g. $R_1R_2C_{sp2}=O_{sp2}$). Here the main problem is deciding the measure of the amount of deformation. In most cases the improper torsion angle is used[19] (figure 2.8).

Figure 2.8 Improper Torsion Angle



As the deformation is likely to be very small it is possible to use a simple harmonic potential energy function so that:

$$E_{oopb} = \sum_{\chi} \tfrac{1}{2} k_{\chi} (180 - |\chi|)^2$$

Where k_χ is the force constant for the out-of-plane bending and χ is the improper torsion angle that will be 180° when the conformation is planar.

2.3.6 Cross Terms

There are several cross terms that can be used to simulate the interaction between different terms in the force field. The main ones used are the stretch-bend, torsion-stretch and bend-bend interactions. These terms can be included, if needed, to create a force field that is as complete as possible and are especially useful when the force field is being used to try to reproduce information such as vibrational frequencies. Ideally it would be advantageous to have all these terms in a force field. However, in the current situation where a force field is required for a large number of atom types without the need to reproduce infra-red vibrational frequencies then it becomes necessary to examine these terms and decide if the increase in accuracy is justified when considering the extra difficulty in optimising the force field.

The stretch-bend interaction is applied in an attempt to simulate the observed situation where the bond length is affected by large variations in the bond angle. If the angle is reduced the bond is found to stretch and similarly its length is reduced when the bond angle is increased. The equation used for this sort of interaction in MM3[8] is:

$$E_{s\theta} = 2.51118K_{s\theta}[(l-l_0)+(l'-l'_\theta)](\theta-\theta_0)$$

Where 1 and 1' are the bond lengths of the bonds with a common central atom with angle θ.

The torsion-stretch interaction is used to simulate the bond stretching that occurs when bonds are eclipsed. The form of this equation from MM3[8] is:

$$E_{\omega s} = 11.995\left(\frac{K_{\omega s}}{2}\right)(l-l_0)(1+\cos 3\omega)$$

The bend-bend interaction is used mainly to help reproduce the correct frequency separation found in spectroscopic investigations of bonds with a common central atom, MM3[8] uses an equation of the form:

$$E_{\theta\theta'} = -0.021914K_{\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0)$$

Where θ and θ' are two valence angles on the same atom and θ_0 and θ'_0 are their appropriate strain free bond angles as used in the respective equations for the calculation of simple valence bond bend energy.

2.4 Force Field Parameters

Deciding on the form of the force field is only the first part of the process. Next, the level at which the parameters should be specified must be decided.

For example, take the strain free bond angle θ_O, which can be specified at many different levels of complexity.

At the simplest level it can be defined so as to have a value relating only to the central atom type (e.g. C_{sp3} 110.5°[20]). The next level of complexity is to define an appropriate strain free bond angle for each of the possible bond angles where the types of the three atoms involved in the angle are used (e.g. C_{sp3}-C_{sp3}-C_{sp3} 109.5°[21]). Finally it is possible, to some extent, to take into account the other atoms connected to the central atom by having a different reference angle for each possible degree of substitution (e.g. C_{sp3} with 3 Hydrogen's H-C_{sp3}-H 108.2°[22]).

It can be seen that this will give rise to a large variation in the number of possible parameters and the same is also true of most of the other parameter types. For example, the torsional barrier (or barriers) can be specified for all four atom types involved or by taking just the central two atoms, a situation that will require far fewer parameters. It is found, as might be expected, that the greater the complexity and number of parameters, the more accurate the optimised force field becomes. This greater accuracy is, of course, made at the expense of the ease and speed of the force field optimisation.

To examine this balance further it is necessary to inspect the numbers involved. To try to optimise a full MM3[8] type force field for the 30 atom types or so which would be needed to cover a reasonable cross section of organic molecules would require:

$\approx$ 300 parameters for each of k_l,l_O,A and B;

$\approx$ 2000 values for $k\theta$,$k\theta^1$ and θ_O;

$\approx$ 10,000 values for V_n.

This is an impossibly large force field to optimise as not only would each parameter have to be optimised in turn but also it would be necessary to make sure that there were a significant number of molecules in the optimising structure set that used each of the possible parameters. That is, to get a reasonable optimisation for a parameter it is necessary to possess information on a sufficient number of cases where that parameter will be used. So if very specific conditions for the application of a certain parameter are set, it can be expected that the parameter will be employed less often than a less specific parameter. Thus there will be a need for more structures if each parameter is to be utilised a similar number of times compared to the less specific case.

If such a large force field is impractical to optimise then a way must be found of reducing the number of parameters that need to be optimised. The obvious way of doing this is to use a lower degree of complexity for the parameters, for example, as stated before, reducing the specificity of the torsional barrier to just specifying the types of the central two atoms will have a drastic effect on the number of parameters required. A problem with this approach is that in going from the most complex but unimplementable levels to the less complex levels there is often an unacceptable drop in the accuracy of the results produced by the force field. There is no point in getting the number of parameters that need optimising down to a workable total if it is then impossible to produce any reasonable results using them.

One possible method for overcoming this problem is that if a pattern can be observed in the parameter values it may be possible to come up with some general rules from which one can calculate the parameters from a combination of some of the already present parameters and a few new ones. For example, calculation of reasonable bond stretching force constant can be done using the bond length parameters and a few general parameters that can be used for all possible combinations of bonds[11].

This approach then gives a way of producing a force field using an intermediate number of parameters between the numbers needed for the various fixed levels of complexity. This approach to producing viable calculated parameters is

covered in greater depth in chapter 4.

Thus the development of a force field can be considered to be in two parts. Firstly the form of the force field must be decided and then, once this is done, it must be decided whether the parameters should be calculated and, if not, then at what level of complexity should they be set at. Only then can an attempt be made to obtain a force field that will reproduce, and also hopefully predict, the structure of the appropriate molecules with the required degree of accuracy.

2.5 Minimisation

As well as producing a value for the energy of a system a method is also required to alter the co-ordinates until an equilibrium situation is reached. This occurs when the total force is zero for each co-ordinate. In most cases this means that the molecule will be in local energy minimum but it should be noted that the condition is also true at an energy maximum or a saddle point. The force is given by minus the partial derivative of the steric energy V_S with respect to each co-ordinate and each component of this should be zero, i.e.:

$$\partial V_S(x)/\partial x_i = 0 \quad i = 1,3N$$

Where N is the number of atoms in the molecule and x is a co-ordinate vector.

The aim of any minimisation technique must be to systematically alter the positions of the atoms until such a situation is reached and that such a system should be both consistent and reliable in its results.

There are several techniques for minimisation in such situations, all of which have advantages and disadvantages, mostly relating to the amount of deformation from which convergence is possible and to the rate at which such minimisation occurs.

Assuming that the energy function is at a point x_S close to the minimum then it is possible to expand the steric energy in a Taylor Series. As x_S is close to the minimum it will be reasonable to truncate the series after the linear terms.

$$V_s(x_s + \partial x) = V_s(x_s) + \sum_{i=1}^{3N} \frac{\partial V_s(x_s)}{\partial x_i} \cdot x_i$$

Taking the case where the value of ∂x is such that $x_S + \partial x$ is the location of the minimum then at this minimum the derivatives of the energy with respect to each of the 3N co-ordinates must be zero. Thus differentiating:

$$\sum_{j=1}^{3N} \frac{\partial V_s(x_s + \partial x)}{\partial x_j} = \sum_{j=1}^{3N} \frac{\partial V_s(x_s)}{\partial x_j} + \sum_{J=1}^{3N}\sum_{i=1}^{3N} \frac{\partial^2 V_s(x_s)}{\partial x_i \partial x_j} = 0$$

This can be shown simply in matrix notation as:

$$\Delta V_S(x_S + \partial x) = \Delta V_S(x_S) + F_S\, \partial x = 0$$

Where ΔV_S is the gradient of V_S, and F_S is the matrix containing the second derivatives of V_S. By subtracting $\Delta V_S(x_S)$ from both sides and then multiplying by the inverse matrix of F_S the following equation can be obtained:

$$\partial x = -F_S^{-1}.\Delta V_S(x_S)$$

Where F^{-1} is the inverse of F.

This approach is used in all the techniques, it is in the way that the various minimisation techniques use this equation that they differ.

2.5.1 Steepest Descent

Witberg[23] used the steepest descent method in 1965, thus making it one of the first minimisation methods used in molecular mechanics calculations. It involves initially shifting each co-ordinate in the molecule by a small increment and calculating the corresponding energy. After the energy has been calculated for each shift the co-ordinate is returned to its initial position. Once all the energies have been calculated each co-ordinate is moved in the direction of lower energy by an amount proportional to the gradient in the steric energy for that co-ordinate. That is to say the co-ordinate will be moved further if it is on a steeper part of the potential energy surface, i.e. :

$$\partial x_i = - L\, \partial V_S(x)/\partial x_i$$

where x is still x_S, our point close to the minimum, and L is a scaling constant.

The constant L is calculated to give a certain root-mean-square change in the Cartesian co-ordinates (σ_0).

$$L = \sigma_0 \left(\frac{1}{3N} \sum_{i=1}^{3N} \left(\frac{\partial V}{\partial x_i} \right)^2 \right)^{-1/2}$$

To get the best convergence it is necessary to change the value of σ_0 during the course of the minimisation process, as a smaller root-mean-square change will be required when approaching the minimum position that when the structure is further away.

The advantages of this system are that it converges well when the geometry, and thus the energy, is far removed from its minimum and that it avoids the chance of getting stuck on a saddle point on the potential energy hypersurface. Its main disadvantage is that when the molecule is nearer to its minimum energy position convergence becomes less efficient because of its dependence on the empirical nature

of L.

2.5.2 Pattern Search

This is a variation on the steepest decent method and was first applied by Schleyer[24]. It uses information from previous iterations to help accelerate movement down valleys in the potential energy hypersurface.

What happens is that at each atom the direction of steepest desent is again found, but in this case a shift is applied to the atom before going on to the next. For each atom both the direction and size of the shifts are preserved. In the next iteration the direction of steepest desent is found once more and if its direction is the same as that atom's previous shift then the actual shift of that atom's co-ordinates is the sum of the steepest descent shift and the last stored shift. In this way a pattern is built up allowing rapid movement down the valleys of the hypersurface. If the direction of the shift changes significantly then the old pattern is abandoned and another one started in the new direction.

This is a better technique but once again requires that an empirical scaling factor be specified for the steepest decent shifts. This scaling factor is critical to the rate of convergence, as with too small a shift the convergence will be unnecessarily slow, but using a larger shift could lead to unpredictability and will tend to result in a higher value for the final energy.

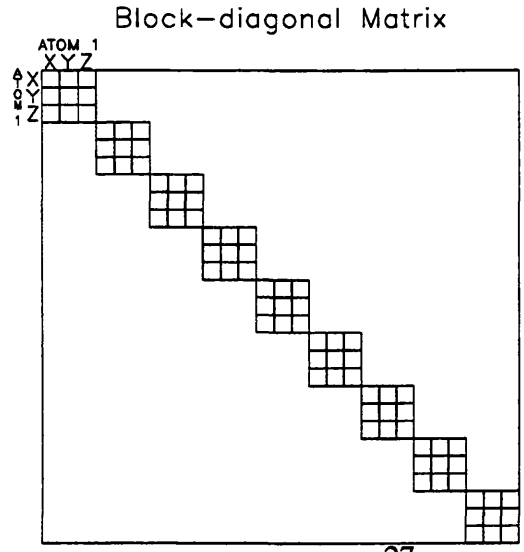
2.5.3 Newton - Raphson

In the previous methods an empirical scaling factor has been used to obtain the shift from the gradient. This has two problems, the first being, it assumes that the slope is flat (i.e. it has a second derivative of zero) and, secondly, that the aforementioned scaling factor has to be specified empirically. The Newton Raphson method overcomes these problems by explicitly calculating the elements of the matrix F. This being the matrix of second derivatives of V_S, i.e:

$$F = \frac{\partial^2 V_s}{\partial x_i \partial x_j}$$

In the full matrix[25,26], a derivative for each possible combination of co-ordinates would be calculated, but, the co-ordinates that will produce the greatest effect on each other are those relating to the same atom. If just the interaction of these co-ordinates are used the result will be a block diagonal[27] matrix F (figure 2.9). One problem with this method is that it is less tolerant to an unfavourable starting geometry.

Figure 2.9 Block Diagonal Newton-Raphson



Another approximation is the pure diagonal[27] method where only the values of $F = \partial^2 V_S / \partial x_i^2$ for each co-ordinate is used, i.e. all off diagonal elements of F are ignored. Both of these methods will minimise well, such that after 50 iterations on a reasonable trial structure it would be expected to get the first derivatives of potential energy with respect to the Cartesian co-ordinates to be down below the level of 0.1 Kcal. mol^{-1} $Å^{-1}$.

If a smaller average force than this is required then it would be necessary to go to a full-matrix Newton Raphson. This involves not only the calculation of the full matrix but also that the matrix should be inverted. This requires significant computation and as such is often only used as a second stage after one of the other methods has bought the structure close to the minimum.

2.5.4 Calculation of Derivatives

The derivatives required can be determined in two ways. In the simplest of these, numerical derivatives, the atomic co-ordinates are moved by a small amount and the energy re-calculated. These energies can then be used to estimate the required derivatives. The second method, analytical derivatives, are determined by applying calculus to the various steric energy terms.

This second method has the advantage that the minimisation is faster as it does not require the multiple energy calculations of the numerical derivatives. However, in a situation where the form of the force field is constantly being changed the numerical derivatives are more useful as there is, in this case, no need to know the form of the force field. That is, when using numerical derivatives only the form of the force field needs to be changed, but with analytical derivatives it would be necessary to derive the new expressions for the changed elements and then use these

to replace the appropriate sections in the program. For analytical derivatives the following sum of derivatives is required:

$$\sum_{i=1}^{3N} \frac{\partial E}{\partial x_i}$$

Each derivative can then be expanded to include all the contributions towards the total energy, i.e. :

$$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial l} \bullet \frac{\partial l}{\partial x} + \frac{\partial E}{\partial \theta} \bullet \frac{\partial \theta}{\partial x} + \frac{\partial E}{\partial \omega} \bullet \frac{\partial \omega}{\partial x} + \frac{\partial E}{\partial r} \bullet \frac{\partial r}{\partial x} + \frac{\partial E}{\partial \chi} \bullet \frac{\partial \chi}{\partial x}$$

The following are typical analytical derivatives with the original formulae from which they were determined.

Bond Length

If the energy for each bond is given by

$$E_l = \tfrac{1}{2} k_l (l - l_0)^2$$
$$= \tfrac{1}{2} k_l (l^2 - l l_0 + l_0^2)$$

then it's derivative will be

$$\frac{dE_l}{dl} = \tfrac{1}{2} k_l (2l - 2l_0)$$
$$= k_l (l - l_0)$$

Angle Bend

If the energy for each angle is given by

$$E_\theta = \tfrac{1}{2} k_\theta (\Delta\theta^2 - k'_\theta (|\Delta\theta^3| - 0.0004 |\Delta\theta^5|))$$
$$\Delta\theta = (\theta - \theta_0)$$

then its derivative will be

$$\frac{dE_\theta}{d(\Delta\theta)} = \tfrac{1}{2} k_\theta (2\Delta\theta - S k'_\theta (3\Delta\theta^2 - 0.002\Delta\theta^4))$$
$$S = \text{sign of } \Delta\theta$$
$$\frac{d(\Delta\theta)}{d\theta} = 1$$
$$\frac{dE_\theta}{d\theta} = \frac{dE_\theta}{d(\Delta\theta)} \frac{d(\Delta\theta)}{d\theta}$$

Torsional Twist

If the energy for each torsion angle is given by

$$E_\omega = V_1(1+s\cos\omega)+V_n(1+s\cos n\omega)$$

then its derivative will be

$$\frac{dE_\omega}{d\omega} = -V_1(s\sin\omega)-V_n(ns\sin n\omega)$$

Out of Plane Bending

If the energy for an out of plane bending angle is given by

$$E_\chi = \tfrac{1}{2}k_\chi(180°-|\chi|)^2$$

then its derivative will be

$$\frac{dE_\chi}{d\chi} = Sk_\chi(180°-|\chi|)$$

$S =$ sign of χ

Non-Bonded Interaction

If the energy for each non-bonded interaction is given by

$$E_r = Ar^{-6} + Br^{-12}$$

The its derivative will be

$$\frac{dE_r}{dr} = -6Ar^{-7} - 12Br^{-13}$$

Coulombic Function

If the energy for each coulombic interaction is given by

$$E_q = k_q\frac{q_iq_j}{r^2}$$

then its derivative will be

$$\frac{dE_q}{dr} = -2k_q\frac{q_iq_j}{r^3}$$

As well as calculating the absolute values of the individual forces it is also necessary to determine the direction of each force so that the overall force on the molecule can be determined. In the case of the Newton-Raphson method the second derivatives need also to be calculated.

For numerical derivatives the equations used are as follows:

First Derivative

$$\frac{dE_s}{dx_i} = \frac{E_s(x_i + \partial) - E_s(x_i - \partial)}{2\partial}$$

Second Derivative

$$\frac{d^2E_s}{dx_i^2} = \frac{E_s(x_i + \partial) + E_s(x_i - \partial) - 2E_s(x_i)}{\partial^2}$$

$$\frac{d^2E_s}{dx_i dx_j} = \frac{E_s(x_i + \partial, x_j + \partial) - E_s(x_i + \partial, x_j) - E_s(x_i, x_j + \partial) + E_s(x_i, x_j)}{\partial^2}$$

2.5.5 Reducing Oscillation During Minimisation

It is found in some situations that have a bad initial structure that from one iteration to the next the root mean square (RMS) force will increase rather than decrease. The reason for this is often that the initial situation has two or more atoms whose van der Waals surfaces are slightly overlapping. This produces a large gradient on the energy surface and causes the atoms to be moved an accordingly large distance apart. In many cases this can move one or more atoms such that the result is a greater overlap with another atom, which will lead to a steeper gradient and so the cycle will continue resulting in a larger effect every time.

When the shift on an atom is calculated it is usually compared with a maximum shift and reduced to this value if it is found to be greater. The reason for this is that a very steep slope would result in the atom moving an unrealistic distance. The technique used to counter the oscillation problem is to allow the program to alter the value of this maximum shift. There are two ways that this has been implemented.

The simplest case is where the energy of the system is calculated at the end of an iteration. This is compared with the lowest energy found so far and, if greater than this by an amount determined empirically from the current maximum shift, the co-ordinates are reset to those at the beginning of the iteration. The maximum shift is reduced and the minimisation is repeated again until either the energy gain is not more than the empirical level or the energy decreases. The main disadvantage with this system is that when using numerical derivatives it increases the time per iteration by about 20%. A major advantage is that the energy of the system is available at the end of each iteration that provides a far more user-friendly indication of the progress of the minimisation than a record of the RMS force.

The second approach uses the fact that it is possible to get an estimate of the energy by summing up the individual energies while going through the atoms in the

molecule calculating the numerical derivatives. This has the advantage that it requires little extra calculation. The draw back is that if the energy of the molecule does start to rise it will tend to appear in the iteration after the bad shift had been made. Thus, if this system is used, the program is required to save the last two sets of atomic positions and return to the earlier of these if any substantial energy rise occurs.

While both these systems have been found to work well, especially in situations involving separate highly charged molecules, the first is preferred because in systems that do require this option there is little time advantage over the second technique and this is more than compensated for by the advantage of being able to see exactly how the energy is varying during the course of the calculation.

Currently this method is only used in the stand alone CHEMMIN minimiser as all the molecules used for the force field optimisation start in configurations that do not contain overlaps.

2.6 References

1 D.H.Andrews, *Phys. Rev.*, 1930, **36**, 544.

2 I.Dostrovsky, E.D.Hughes and C.K.Ingold, *J. Chem. Soc.*, 1946, 173.

3 T.L.Hill, *J.Chem. Phys.*, 1946, **14**, 465.

4 F.H.Xestheimer and J.E.Mayer, *J. Phys. Chem.*, 1946, **14**, 733.

5 D.H.R.Barton, *Experimentia*, 1950, **6**, 316.

6 J.A.Pople and D.L.Beveridge, "Approximate Molecular Orbital Theory", McGraw-Hill, New York, 1970.

7 A.T.Hagler, S.T.Stern, S.Lifson and S.Ariel, *J. Am. Chem. Soc.*, 1979, **101**, 813.

8 N.L.Allinger, Y.H.Young, L.Jenn-Huei, *J. Am. Chem. Soc.*, 1989, **111**, 8551.

9 D.H.Wertz and N.L.Allinger, *Tetrahedron*, 1974, **30**, 1579.

10 E.A.Mason and M.M.Kreevoy, *J. Am. Chem. Soc.*, 1955, **77**, 5808.

11 D.N.J.White, J.N.Ruddock and P.R.Edgington, *Molecular Simulation*, 1989, **3**, 71.

12 N.L.Allinger, *J. Am. Chem. Soc.*, 1977, **99**, 8127.

13 J.M.Lehn and G.Ourisson, *Bull Soc. Chim. Fr.*, 1963, 1113.

14 J.E.Lennard-Jones, *Proc. R. Soc. London, Ser A*, 1924, **106**, 463.

15 A.D.Buckingham and B.D.Utting, *Ann. Rev. Phys. Chem.*, 1970, **21**, 287.

16 H.Margenau and N.R.Kestner, "Theory of Intermolecular Forces", Pergamon, Oxford, 1969.

17 D.E.Williams, *J. Chem. Phys.*, 1965, **43**, 4424.

18 D.E.Williams and T.L.Starr, *Comput. Chem.*, 1977, **1**, 173.

19 A.Warshel, M.Levitt and S.Lifson, *J. Mol. Spectroscopy*, 1970, **33**, 84.
20 E.K.Davies and N.W.Murrall, *Comput. Chem.*, 1989, **2**, 149.
21 J.G.Vinter, A.Davis and M.R.Saunders, *J. Comput.-Aided Mol. Design.*, 1987, **1**, 31.
22 D.N.J.White and M.J.Bovill, *J. Chem. Soc. Perkins*, 1977, **II**, 1610.
23 K.B.Witberg, *J. Amer. Chem. Soc.*, 1965, **87**, 1070.
24 E.M.Engler, J.D.Andose and P. von R. Schleyer, *J. Amer. Chem. Soc.*, 1973, **95**, 8005.
25 S.Lifson and A.Warshel, *J. Chem. Phys.*, 1968, **49**, 5116.
26 E.J.Jacob, H.B.Thompson and L.S.Bartell, *J. Chem. Phys.*, 1967, **47**, 3736.
27 D.N.J.White and G.A.Sim, *Tetrahedron*, 1973, **29**, 3933.

CHAPTER 3

Parallel Computers and occam

3.1 Introduction

Until recently computer programs could only be executed sequentially, that is to say, that a single central processing unit (CPU) would extract an instruction from memory, which it would then execute, before going on to execute the next instruction. This is though, a very artificial method of operation and in the human brain we have a good example of how effectively executing many computations at once allows tasks to be performed in an instant which, even the most powerful computers, have found great difficulty in doing, for example, object and speech recognition.

This method of having many processors working together at the same time is called parallel processing. The underlying concept is that rather than trying to make any particular processor run ten times faster the task is divided up so that, for example, ten processors can each do one tenth of the required calculation, resulting in a ten fold speed-up when compared to a single processor.

The first computers that could be said to be in any way parallel were vector processors, these divided up operations on arrays between many processors. This is usually described as fine grain parallelism as the parallelisation occurs at a low level. Another type of fine grain parallelism is where each parallel task is very small. One of the first computer to demonstrate this type of parallelism was the DAP (Distributed Array Processor) which was originally produced by ICL but is now produced by AMT. This initially consisted of a 64x64 array of simple single processors. These processors would then be given a 64x64 array of data on which to work. All the processors executed the same code and so the DAP is described as a SIMD (Single Instruction Multiple Data) machine. The ICL DAP had several drawbacks, including the need for an ICL Mainframe host and the lack of a reasonable computer graphics interface. The AMT DAP has overcome these problems and has made the DAP a more viable proposition for tasks open to the SIMD approach[1].

At about this time the first laboratory-sized parallel computer systems became available. These consist of a collection of 16 or 32-bit microprocessors that are multiply interconnected. Each processor is part of a node that also contains a floating point unit (FPU), 128 Kbytes or more of memory and a number of channels to connect with the other nodes. Some of these machines attempt to take the parallelism further by incorporating a vector processor on each node, these systems are usually referred to as 'turbo' systems. The node connections are often fixed in a hypercube configuration. Examples of this sort of machine include the Intel iPSC Hypercubes, the Alliant FX series and the FPS T series (an example of the 'turbo' variety of machine[2]). Unlike the DAP these machines can be set up with different code running

on each node, so rather than the SIMD of the DAP these are described as MIMD (Multiple Instruction Multiple Data) machines.

The fixed topology and the need in the 'turbo' case to both parallelise and then vectorise means that these machines are difficult to program, especially if maximum efficiency using all nodes is required.

This was the state of parallel processing, until, in 1985 the Inmos company produced the first versions of the transputer - a computer on a chip, which was specifically designed from the very start to be used as part of a parallel processing engine[3]. As well as producing the hardware that can handle programs running in parallel they also saw the need for an associated language that would support parallelism as a main construct of the language. The result of this is the language 'occam'[4], which though implementable on other machines has a special relationship with the transputer series.

A feature of occam on the transputer is that it is easy to test parallel programs on a single processor before running the code on a full network.

With the transputer and occam it is now possible for the first time to get supercomputer performance within a reasonable price range and it is this power that has made the method of parameter optimisation here-in described possible.

3.2 The Transputer

There are three points that make the transputer, and the high performance T800 variant in particular, ideally suited for parallel processing.

The first of these is the occam programming language that will be described in detail later in this chapter. Secondly, even when looked at as a single processor, the transputer was more powerful than most other chips on the market at the time. This is especially true of the T800 version with its on-chip maths co-processor. Finally, its inter-processor links allow for the construction of many different network configurations and mean that the transfer of data around any such network is a relatively simple process.

There are several points to note about the T800 transputer (Figure 3.1). Unlike many other systems at the time that use floating point co-processors the transputer has the floating point unit (FPU) as part of the silicon on the main chip rather than as a separate chip. This makes the communication between the CPU and the FPU both easier and faster. There is also 4K of internal RAM which means that with careful programming, the stack and some of the more often called subroutines can be placed in the internal RAM and so can be accessed and run far faster than if they had been placed in the external memory. It also means that in some cases, such as a graphics pipeline, there may be no need for any expensive external memory to be

used at all. The third point of interest concerns the four links. These are the high speed data channels that can be used to connect processors and pass data very simply between them at up to 10Mbits/sec. These links also run in parallel with the CPU so a program performing communications and calculations in parallel will actually run faster than if these tasks were being done sequentially. Both the CPU and the FPU use what is known as the RISC approach (**Reduced Instruction Set Chips**). This means they are optimised to carry out a few instructions rather than having a large and complex set of possible instructions.

Figure 3.1 The T800 transputer

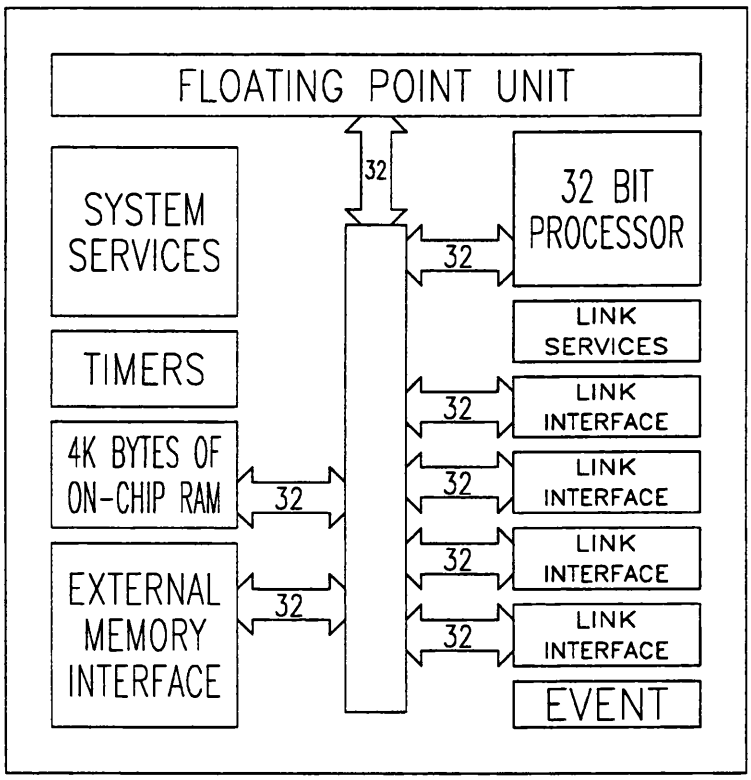


Table 3.1 Comparison of processor computing power

Processor	Clock speed	Whetstones/sec
IMS T800-20	20Mhz	4600K
80386/WTL 1167	16Mhz	3600K
Fairchild Clipper	33Mhz	2220K
Intel 80386/80387	16Mhz	1354K
VAX-11/780 FPU	n/a	1028K
MC 68020/68881	16.67Mhz	1028K
NS 32332/32081	15Mhz	728K
IMS T414-20	20Mhz	663K
Intel 80286/287	8Mhz	300K

Table 3.1 gives an idea of the raw processing power of the T800 transputer as compared with other processors available at the time. The Whetstone benchmark uses data processing and floating point operations and gives a good idea of the computing power available for scientific calculations. It should be noted that there was soon available a 30Mhz version of the T800 processor with a commensurate increase in the computing power.

3.3 Networks

Another difference between normal processing and parallel processing is that with parallel processing it is necessary to consider for each application the optimum arrangement of the processors in a network. That is, connecting the links in different ways will produce different network topologies, which will be suitable for different tasks.

Figure 3.2 Typical network topologies

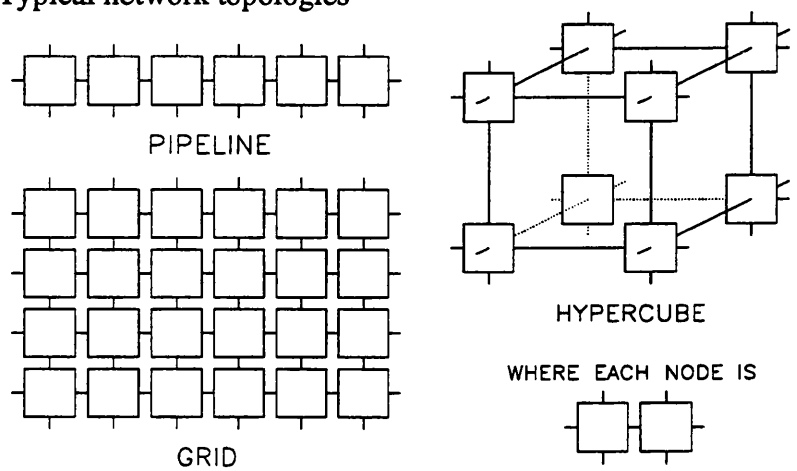


Figure 3.2 shows some typical networks. The pipeline could be used for situations such as graphics, where each processor would do a different task such as scaling and shading. The grid would be more appropriate for programs doing some sort of matrix calculation to allow easy mapping of the problem onto the network. The hypercube could be used in situations where the need is for the shortest communication distance between arbitrary processors. However, this configuration is currently impossible to implement with single processors as transputers do not, currently, have the 6 links each that this configuration requires.

Special switching chips can be used so that the configuration of the network can be set up from software rather than having to be explicitly hard wired to a new configuration in each case.

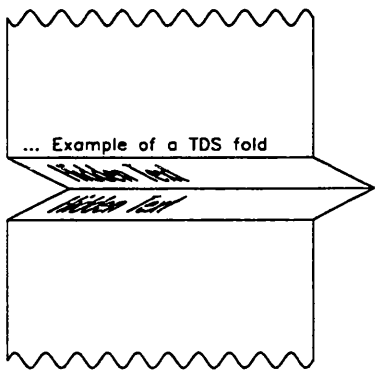
3.4 The TDS

TDS[5] stands for the Inmos Transputer Development System, which is an

environment running both on a transputer and on its host machine such as a microvax or an IBM PC. From within the TDS all necessary program development can be achieved. The distinguishing concept used by the TDS is that of folds.

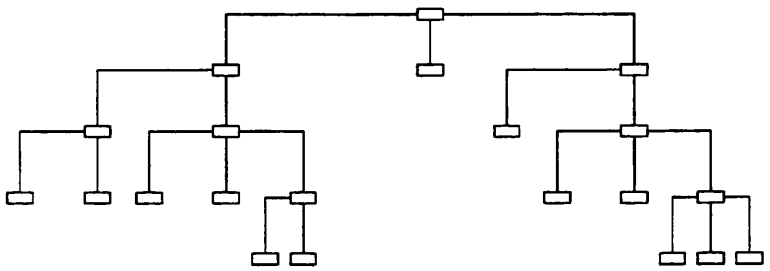
If you consider the program as written on a sheet of paper then the paper can be folded such that some of the code is hidden in a fold and that the place where the code is hidden is shown by the crease, which in the case of the TDS can be labelled to show what the fold contains (figure 3.3).

Figure 3.3 A TDS Fold



Folds can be nested to any depth with multiple folds possible within any one fold. This allows programs to be developed in a top down approach, as a program can be broken down into tasks each in its own fold, each of which is then broken down into more tasks and folds until the tasks get to the size where they can be easily programmed (figure 3.4). This use of the fold structure leads to more readable code as it is possible to produce most folds such that they can fit on the screen in their entirety.

Figure 3.4 Task Division



However with the advantages that come with the fold system also comes the disadvantage of moving between sections of the code that must be done by use of the cursor keys. So to travel from one location to another may involve going up many levels and then back down a different branch in the fold tree. Going up a fold tree is simple as it just requires a single key press to exit from a fold. However, when going down folds the cursor has first to be placed on the required fold in each case before it

can be entered. This was not made any easier when the TDS was upgraded and the position at which the cursor appears in a new fold changed!

The TDS takes the concept of folds further in that everything, including code, data and utilities, is contained within folds and operations are then performed on these folds. Thus the compiler is obtained by extracting it from its fold and it can subsequently be applied to a program fold with the resultant production of a code fold that can then be extracted and run.

Within the TDS environment access is available to all the required system utilities, the main ones being the occam syntax checker, compiler, linker and configurer. In the later editions of the TDS there is a debugger with which it is possible to trace the position at which a fatal run-time error occurred, obtain information as to the values of the array elements, retrace back up through the procedure calls, or find where the program is waiting on links.

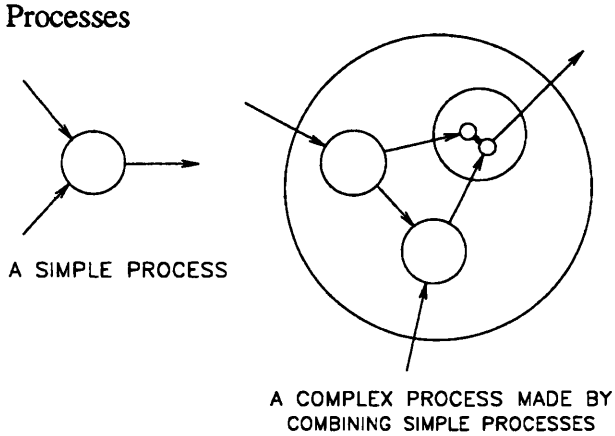
The TDS also includes utilities for writing to and from DOS files and for doing standard editing functions, such as search and replace functions. Other included functions such as move and copy fold are specific to the TDS and provide a quick and easy method for copying and transferring any section of code.

3.5 Occam

Occam is a language based on the idea of processes. A process is an independent computation that can communicate with other processes being executed at the same time (figure 3.5). Occam thus allows the building up of complex concurrent systems by connecting various simpler processes.

As with the top down approach used in the folding editor of the TDS, occam is a language where complicated processes can be made up from very low level primitive processes. A comparison can be also drawn with the transputer RISC architecture, in that both have a limited number of fast instructions, rather than having a large selection of complicated instructions.

Figure 3.5 Occam Processes



The main primitive processes in occam are :-

`v := e`	assigns the value of expression e to the variable v
`c ! e`	sends the value of the expression e down channel c
`c ? v`	Reads in the value v from channel c

Example of these are:

```
co.ordinate[0][i] := co.ordinate[0][i] + offset
input.channel ! new.value[i][j]
output.channel ? result
```

These are combined to form constructs. All constructs must start with one of the following which tells the compiler how the processes in the construct are to be implemented:

3.5.1 SEQ

Processes following this directive are executed sequentially, for example:

```
SEQ
  in.chan ? var1
  var1 := var1 + var1
  out.chan ! var1
```

This section of code reads in the value of var1 from the channel in.chan, doubles its value and then sends it down the out.chan channel. It can be seen that a sequential process works in a similar way to a program written in a more conventional language, in that it starts, executes the commands in order and then terminates.

In occam the double indenting of the commands is critical as it informs the occam compiler which of the following processes are part of the current construct. It is also always necessary to specify the type of construct for any situation where there are two or more processes. This is because, unlike in other languages, the processes cannot be assumed to run sequentially. Thus it is necessary to tell the compiler how each particular construct is to be run.

Occam also allows processes to be replicated, which for the SEQ construct results in something akin to a normal indexed loop in FORTRAN, for example:

```
SEQ i = 0 FOR 10
  channel.out ! i
```

This code would output the numbers 0 to 9 down the channel.out channel. This bring to light another difference between occam and FORTRAN in that with FORTRAN the default base number is 1, i.e. the arrays all start at the first element, in occam

though this value is zero. This can cause problems if processes in the different languages need to communicate. For example, what is considered as the 8th atom in the FORTRAN will have an array index of 7 in the occam as element 0 refers to atom 1.

3.5.2 WHILE

The WHILE command is not really a construct directive. It is however, placed before a process or construct and allows it to be executed repeatedly, until the condition following the WHILE command ceases to be true. Thus it can be considered to be similar to a DO-UNTIL loop, for example:

```
WHILE x <> 0
  SEQ
  chan.in ? x
    chan.out ! x
```

In this case the process will continue to read values of the variable x from chan.in and send them down chan.out until the value of x is zero. When the zero value is received it will be sent down the out.chan as usual after which the construct will terminate. If the condition is given as TRUE rather than as a test, then the loop will operate indefinitely.

3.5.3 PAR

The processes following the PAR construct directive are executed in parallel. This construct will only terminate when all the components of the construct have terminated, for example:

```
PAR
  SEQ
    chan1 ? x
    chan2 ! x
  SEQ
    chan3 ? y
    chan4 ! y
```

In this case the construct will only terminate when the value of x has been passed from chan1 to chan2 and y has been passed from chan3 to chan4.

Parallel processes should only communicate between themselves using channels and each channel should only be used in two processes at any one time. Also, if the compiler checks are to be strictly applied then each process should operate on its own independent set of variables. For example, the following would cause an error during compilation if the variable x is declared outside the PAR

construct:

```
PAR
  SEQ
    chan1 ? x
    chan2 ! x
  SEQ
    chan3 ? x
    chan4 ! x
```

As with the SEQ case, there is also a replicated PAR. In this case the replication allows a set number of identical processes to be set running at the same time. For example:

```
PAR i = 0 FOR 9
  WHILE TRUE
    SEQ
      buffer[i] ? b.data
      buffer[i+1] ! b.data
```

As part of a larger program this section would act as a buffer reading in data, then passing it through the buffer channels until the values are extracted out of the buffer[10] channel. For this to work properly it would be necessary to declare the b.data variable after the PAR so that each of the parallel processes worked with a different b.data variable. This example also shows another feature of occam in that channels can be set up in arrays, this is explained in more detail in the later section on declarations.

3.5.4 IF

This will perform a number of tests in order, executing the first process where the conditions are met. Care must be taken to make sure that one of the conditions will be met as a program will hang if all the various tests fail. It is usual practice, unless totally sure that a condition will be met, to finish with a condition that will always be met. For example:

```
IF
  x > 0
    x := x + 1
  x = 0
    x := x + 2
  TRUE
    SKIP
```

The command SKIP is a process in occam that does nothing. At first sight a process which does nothing may seem to be rather strange, but in many situations a process of some sort is required by the syntax where no action is needed. It is just as legitimate to put something which has no effect, such as x := x, but the use of the SKIP command allows a more elegant solution.

A similar process is the STOP command which unlike the SKIP command does not terminate after doing nothing. On the face of it, this may appear even more pointless than the SKIP command. STOP however can be used in error handling, so when an error is found, STOP will halt the process and the relevant point can later be found using the occam debugger.

Again occam allows the use of replicators:

```
IF
  IF i = 0 FOR number.of.elements
    array[i] = 0
      last := i
  TRUE
    last := -1
```

In this case, the array is being searched for the first location containing zero and its location is stored in the variable last. If no zero values are found then the variable last is set to -1.

3.5.5 ALT

ALT stands for alternative construct. This is associated with the synchronised way that messages are passed through the channels. It is possible to imagine a situation with several processes running in parallel where one of the processes has several incoming channels and is required to act differently depending on which channels first receives data. It is in this situation that the ALT command is used.

The ALT command can be compared with the IF command in that both involve the selection of a process from a set of possible processes, where each has an associated test. The IF command decides which process to execute by going through the logical tests in order until one tests true. In the ALT case the tests relate to when data comes down a specific channel. The first channel down which data is successfully received will cause the execution of its associated process. For example:

```
ALT
  chan1 ? x
    x := x + 1
  chan2 ? x
    x := x + 2
  chan3 ? x
    x := x + 3
```

The result of this construct is that the number added to the received value x depends on which channel the value was initially received from.

The ALT construct is very useful for situations where some sort of software interrupt may be required. For example, passing a value down a close-down channel to terminate all the processes in a pipeline.

Replication is again allowed and is used typically to provide an elegant way of checking arrays of channels for incoming data. For example:

```
ALT i = 0 FOR num.of.channels
  chan.array[i] ? x
    array.number := i
```

The result of this code is that the first chan.array to receive data is the first to be acted upon.

It is also possible to combine a logical IF type test into the ALT by adding an optional condition to the guard. This means that not only must a channel be about to receive data, but that the associated condition must also be true. So, for example, if a signal from a channel would cause a process to increment a variable then a condition could be set on the normal guard so that the process would only be executed if the data was about to be sent and the variable was not at its maximum value. The following are examples of allowable guards:

```
(y < 100) & chan1 ? x
(y < 100) & SKIP
```

In the second case rather than get data from a channel a SKIP command has been used, so the associated process does not require any input from a channel and may be executed if y is less than 100. The order in which the guards are checked is random and so even though this condition is always ready if y is less than 100 it may well be that there is another of the guards ready and this other guard may be checked first. So having an always ready guard does not mean that that process is sure to be executed.

There are many situations where it would be advantageous to always check the channels in a specific order and for this case the priority alternative (PRI ALT) construct may be used. For example:

```
WHILE going
  PRI ALT
    stop.chan ? in.val
      going := FALSE
    TRUE & SKIP
      ... else carry on with main loop
```

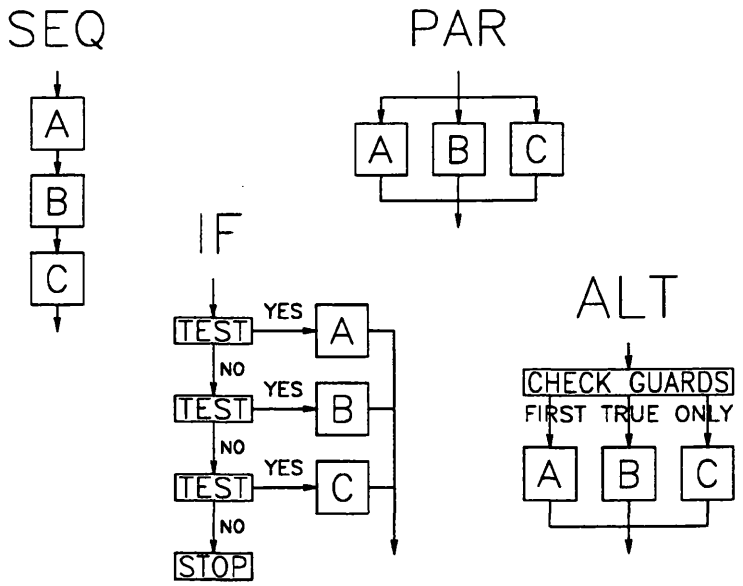
Each time this process is reached the stop.chan channel is checked to see if any data is ready to be received and if so the WHILE process is terminated. If nothing is ready to be transmitted down the stop.chan channel then the second guard is checked, which in this case is always true and so the rest of the process is executed. The three dots (...) indicate that there is a TDS fold at this location.

A similar priority term can be added to the PAR construct. This can be used for things such as high priority message passing. For example:

```
PRI PAR
  WHILE TRUE
    SEQ
      message.channel.in ? message
      SEQ k = 0 FOR bytes.per.message
        message.channel.out ! message[i]
  ... rest of program
```

Figure 3.6 gives a graphical overview of the different construct types.

Figure 3.6 Construct Types



3.5.6 Types and Specifications

All the previous examples are not in fact strict occam as they contain no statements specifying the types of the various variables used. The statements used to declare the types can also be used to specify the size of any arrays used.

All variable names should be in lower case, to distinguish them from keywords, such as PAR, and the names can be of any length. When made up of more than one word the individual words must be separated by dots, so this.is.a.very.very.long.variable is an example of a valid variable name. Each variable must have its type declared. The types available within the current

implementation of occam are:

```
BYTE      -- single byte often used to store characters
INT       -- integer of default size (usually 32 bit)
INT16     -- 16 bit integer
INT32     -- 32 bit integer
INT64     -- 64 bit integer
REAL      -- real of default system size (usually 32 bit)
REAL32    -- 32 bit real
REAL64    -- 64 bit real
BOOL      -- logical, can be either TRUE or FALSE
```

Channels must be declared as type CHAN OF ANY. Some examples of valid declarations are:

```
[mxang]REAL32 angle :
[mxang][4]INT ang.atoms :
BOOL looping :
[number.of.minimisers]CHAN OF ANY to.minimisers :
```

It is possible to set constants. For example:

```
VAL zero32 IS 0.0(REAL32) :
VAL no.char IS 0(BYTE) :
```

It is also possible to abbreviate parts of arrays, either as constants:

```
VAL first.element IS store[1] :
```

or as a variable

```
first.element IS store[1] :
```

The difference is that the variable can be used as a sort of shorthand to access the array element, whereas the constant is just the value of the array element at the time that it was declared.

It is also possible to select arrays out of larger dimensioned arrays. This is very useful when, for example, access is required to the data relating to a single molecule. So if the initial declaration is:

```
[mxmol][mxat]REAL32 charge :
```

then using

```
chargei IS charge[i] :
```

results in an array called chargei of dimension mxat that contains the charges for the molecule i only.

3.5.7 Scope

A very important concept in occam is that of the scope of a variable. All occam declarations are local to the processes that follow their declaration. This range of the variable is called its scope. The scope is determined by the level of indentation at which the variable is declared. For example:

```
BOOL flag1 :
SEQ
  ... section one
  BOOL flag2 :
  SEQ
    ... section two
    ... section three
```

In this case the scope of flag1 is over all three sections, whereas the scope of flag2 covers only the sections two and three.

3.5.8 Operators

The elementary operators in occam are:

`a + b`	--- simple addition
`a - b`	--- subtraction
`a * b`	--- multiplication
`a / b`	--- division
`a \ b`	--- remainder when a is divided by b

In occam there is no priority for the order that operations are executed so in all cases where more than one operation is taking place in a process the order of execution must be specified by the use of brackets. For example:

```
(3+4)*(2+3) = 35      3+(4*(2+3)) = 23
```

Logical values are obtained by the following operations:

=	--- equal
<>	--- not equal to
>	--- greater than
<	--- less than
>=	--- greater or equal to
<=	--- less than or equal to

These are used mainly in WHILE, IF and ALT constructs as tests for required conditions.

3.5.9 Procedures

There are certain situations where it is useful to move code from within a program and place it at the front of the program in the form of a procedure. The main advantage of doing this is that it allows the process to be called from several different locations within the program. This means that only one copy of the code is required rather than having a copy at each point where it is needed. This is obviously advantageous as any change in the process need only be done once. The following is an example of a valid procedure:

```
PROC add.squared(CHAN OF ANY chan.in,chan.out,
                 INT value1)
     INT value2 :
     SEQ
       chan.in ? value2
       value2 := value2 + (value1 * value1)
       chan.out ! value2
:
```

To execute the procedure it is then just a case of calling the procedure name and passing to the procedure any data and channels that it requires. So for the above example, the procedure could be called as follows:

```
add.squared(chan.in,chan.out,value.in)
```

The procedure is terminated with a colon such that it is in line with the beginning of the PROC statement. In general anything terminating in a colon is considered to be associated with any processes that follow it and that are within its scope. So just as the declarations of an integer variable will be attached to the process or processes that follow after its colon, so the PROC will be attached to whatever follows its colon.

Even when a process is only going to be used once within a program it can still be useful to place the process in a procedure at the outermost level. One reason for doing this is that the call to the procedure may be many folds deep into the program but if many alterations to the code of the process are required then they can be done at the uppermost level in the program where the procedure is stored. It also means that any general purpose routine can be converted to a procedure that can then be easily copied from its original program to any others written later which might require the routine.

The following is an example of occam code which uses a procedure. In this

example the process will read values as they arrive on either chan.in.1 or chan.in.2, square these values, add one and then send the result down the chan.out channel. The process will only terminate when it has received the number zero from both of the input channels. Note that this procedure uses all of the above described construct types except PAR.

```
PROC square.plus.one(REAL32 in.value,out.value)
  VAL one IS 1.0(REAL32):
  SEQ
    out.value := in.value * in.value
    out.value := out.value + one
:
CHAN OF ANY chan.in.1,chan.in.2,chan.out :
REAL32 in.value,out.value :
BOOL flag1,flag2,going :
SEQ
  flag1 := TRUE
  flag2 := TRUE
  going := TRUE
  WHILE going
    SEQ
      ALT
        chan.in.1 ? in.value & flag1
          SEQ
            square.plus.one(in.value,out.value)
            IF
              in.value = 0
                flag1 := FALSE
              TRUE
                SKIP
        chan.in.2 ? in.value & flag2
          SEQ
            square.plus.one(in.value,out.value)
              IF
                in.value = 0
                  flag2 := FALSE
                TRUE
                  SKIP
      chan.out ! out.value
      IF
        (.NOT.flag1) AND (.NOT.flag2)
          going := FALSE
        TRUE
          SKIP
```

3.5.10 Libraries

Instead of putting the PROC's before the program it is possible to place them into libraries. It is then possible to include the library at the location that the procedure should be declared by using the term:

```
#USE square
```

As well as putting procedures into libraries it is also possible to create libraries containing other code such as declarations. This can be useful as it allows declarations required throughout a large program to be turned into one or more libraries which can then be used in place of the declarations themselves.

There are also libraries that are supplied with occam. These contain procedures for input/output and for calculation of functions such as sine's, cosine's and square roots. Thus to use the function EXP (exponent) for REAL32 variables the following statement would need to be placed before the required procedure:

```
#USE snglmath
```

3.5.11 Multiprocessor Programs

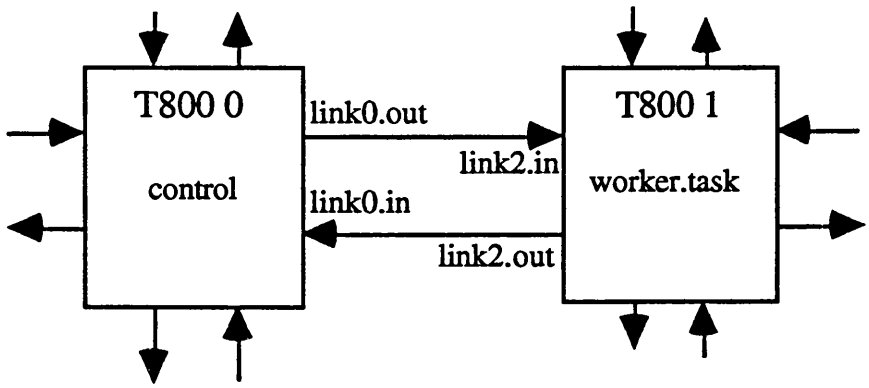
So far the description of occam has not taken into account that eventually an occam program will probably require different processes to be running on different processors. To allow for this there are several alterations to a program that must be done.

First the program's channels must be tied to the inter-transputer links using the PLACE command. Then the PAR statement above the elements to be run on separate processes needs to be replaced by the PLACED PAR command. This is similar to the PAR command, except that it is used to specify which processor each process is to run on. For example, when a situation requires that a control program is to be run on processor 0 and a worker task process run on processor 1 the code could be as follows:

```
... SC control
... SC worker.task
VAL link0.out IS 0 :
VAL link0.in  IS 4 :
VAL link2.out IS 2 :
VAL link2.in  IS 6 :
CHAN c1,c2:
PAR
  PLACED PAR
    processor 0 T8
    PLACE c1 AT link0.out :
    PLACE c2 AT link0.in :
    control(c1,c2)
  PLACED PAR
    processor 1 T8
    PLACE c1 AT link2.in :
    PLACE c2 AT link2.out :
    worker.task(c1,c2)
```

This arrangement of processes of processors is shown graphically in Figure 3.7.

Figure 3.7 Example arrangement for a multiprocessor program



The term SC refers to separately compilable filed folds which contain the code for the control and the worker tasks and so allows, during development, for the sections to be compiled independently. The processor 'in' links are numbered zero to three and the processor 'out' links are similarly numbered four to seven.

The processor statement after the PLACED PAR tells the configurer the number of the transputer and its type. Thus in this example both processor 0 and processor 1 are T800's. This is followed by the PLACE statements and finally the name of the procedure.

The advantage of the PLACED PAR system is that large programs can be written and tested on a single transputer using the normal PAR construct which uses time slicing between the parallel processes and then when the program has been developed and tested it is a simple task to change the configuration to allow the program to run on a network of processors.

3.6 Other Languages

As well as occam there are other languages available for the transputer such as FORTRAN-77, Pascal and C. These, however, are stand-alone compilers where a normal ASCII file in compiled to an object file which is then linked with libraries to form the final executable program. In all these cases it is necessary to include additions to the languages to allow for the use of parallelism and the possibility of programs running concurrently on different processors. The advantage of using these languages is that there is a large body of code already available where the advantage in converting the programs to occam would be small. In many cases the best solution to the problem of whether to use the original code or to rewrite the program in occam is to leave the user interface and the non computationally intensive parts of the

program in the original language but to convert the computationally intensive parts to occam. The latter could then be run on a network, connected to the main processor which runs the remaining parts of the program written in the original language. This is the approach taken with the COMMET molecular modelling system where only the more computationally intensive parts such as molecular mechanics calculations are converted to occam.

3.7 References

1 I.S.Reid and J.K.Reid (Eds.), "Vector and Parallel Processors in Computational Chemistry", North-Holland, Amsterdam, 1985. ISBN 0-444-86794-3.

2 J.L.Gustafson, S.Hawkinson and K.Scott, *J.Parallel & Distributed Computing*, 1988, **3**, 297.

3 M.Homewood, D.May, D.Shepherd and R.Shepherd, *IEEE Micro*, 1987, **7(5)**, 1.

4 a) D.Pountain and D.May, "A Tutorial Introduction to Occam Programming", BSP Professional Books, Oxford, 1987, ISBN 0-632-01847-X.

b) K.C.Bowler, R.D.Kenway, G.S.Pawley and D.Roweth, "An Introduction to OCCAM2 Programming", Chartwell-Bratt, Sweden, 1987, ISBN 0-86238-137-1.

5 D.Cormie, "Getting Started with the TDS", Inmos Technical Note 3, Document No. 72-TCH-003-00, Inmos, Bristol.

CHAPTER 4

Levels of Parameter Specification and Alternative Methods Using Parameter Calculation

4.1 Introduction

As stated in chapter 2, a full MM force field for a reasonable selection of atom types would be almost impossible to implement and optimise because of the large number of parameters that this would require.

Our aim must be, therefore, to produce a force field that gives a reasonable degree of accuracy by using various methods for reducing the number of parameters that are needed, so that the task of optimising these parameters becomes viable.

For the majority of interaction types there exist methods that can be used to reduce the number of parameters required. What is needed is some way of deciding where the compromise position should be in each case. That is, in each case obtain the correct balance of the optimisation benefits gained from the reduction in the number of parameters against the resulting loss in accuracy.

The OPTPARAM program (see chapter 5) has been written to allow the evaluation of various combinations of these methods of reducing parameter numbers. Initially, before examining this program in detail, the methods of parameter reduction must be investigated.

These methods can be divided into two general groups.

The first involves deciding exactly how specific the situation is that distinguishes when any particular parameter should be used. For example, in the angle case, it is possible to use a parameter dependent only on the central atom or on all the atoms involved in the angle or all the atoms in the angle plus the degree of substitution on the central atom. It is obvious that the less specific the conditions relating to when a particular parameter will be used are, the fewer parameters will be required to allow for all of the possible circumstances.

Secondly, there are the cases where, instead of specifying the parameters exactly, they are instead calculated from other data. The data used is usually other parameters, a few general equation constants, usually polynomial coefficients and, in some cases, the connectivity of the molecule under consideration. So, for example, rather than specify the bond stretch force constant for each possible bond, an attempt could be made to find equations with which to produce these required parameters.

Another possible technique is to produce the force field using one of the above approaches but to have a few explicit parameters which will be used instead of the general ones in certain specific situations. An example of this sort of approach is given in the section below on torsion barriers.

The following sections examine each of the interaction types in turn and show for each case what the possible methods for reducing the number of parameters are. In all these cases the methods which involve changing the level of complexity or the

equation describing the potential will be examined first and this is followed by what formulae, if any, are available to calculate the appropriate required parameters.

4.2 Bond Stretch

There are two parameters required for each bond stretch interaction. The reference bond length l_0 is one of the few parameters that is not accessible to reduction in any way and in all current force fields the reference bond lengths are specified exactly for each possible bonding interaction. There are however, several levels possible for the force constant k_l.

The bond lengths between atoms are generally considered to be 'hard' geometric properties. The term 'hardness' means that the property requires a larger amount of energy to deform it compared to a 'softer' property such as torsional angle. As such the bond lengths will usually vary little from their ideal values so at the extremely simple level the bond stretch constant k_l can be infinite for all cases, thus effectively fixing the bond lengths. One drawback to this method is that any minimisation must take place using internal co-ordinates as any Cartesian co-ordinate shift that alters the bond lengths will be heavily penalised. This will be refered to as the 'fixed' option for bond stretch force constant determination.

The next possibility is that the bonds should be allowed to alter, but to give them all the same force constant's dependant only on the bond multiplicity. This means that all single bonds would have one value for the bond stretch force constant and so on. This will be refered to as the 'multiplicity only' option for bond stretch force constant determination.

The final possibility is that there should be a different value for the bond stretch force constant for each reference bond length (i.e. each value of l_0 would have its own optimised value of k_l). This will be refered to as the 'full' option for bond stretch force constant determination.

Investigation into the possible calculation of bond strength force constants from bond lengths had been done by White[1], who by examining the relationship between these parameters produced the following two equations:

$$\tfrac{1}{2}k_l(ij) = \left[\frac{c_1}{l_0(ij)^2}\right] + \left[\frac{c_2}{(l_0(ij)-1)^2}\right] - \left[\frac{c_3}{l_0(ij)}\right]$$

$$\tfrac{1}{2}k_l(iH) = \frac{c_4}{l_0(iH)^2}$$

The first equation refers to all situations, except those where a hydrogen atom is involved, in which case the lower equation should be used. c_1, c_2, c_3 and c_4 are parameters which would initially be calculated from such graphs but could be

optimised in just the same way as any other parameter. This will be refered to as the 'calculated' option for bond stretch force constant determination.

4.3 Angle Bending

The parameters for the angle bending interaction are not only some of the more important parameters, they also have the greatest range of options as to the way that the parameters can be decided upon. This is because the interaction requires values in each case for the bending constant k_θ (and possibly k_θ'and k_θ") and the reference bond angle θ_0 and the methods by which these two types of parameters are determined can be entirely different. For example, the angle bending constant could be selected from a list depending on the atoms in the angle whereas the reference angle itself could be calculated using some general scheme.

At the very simplest level the values of k_θ and θ_0 would depend on just the type of the central atom[2]. For example, all bond angle interactions with C_{sp3} as the central atom would use the same values of k_θ and θ_0. This will be refered to as the 'central only' option for determining angle related parameters.

There is then a sizeable jump in the number of parameters when going to the next situation where both k_θ and θ_0 can be specified depending upon the three atom types which make up the selected angle[3]. So that in this case C_{sp3}-C_{sp3}-C_{sp3} would have different values for k_θ or θ_0 compared to the angle C_{sp3}-C_{sp3}-H. This will be refered to as the 'full' option for determining angle related parameters.

The final complexity level is usually only used for the θ_0 term and is where the degree of substitution at the central atom is taken into account[4]. For example, the values of θ_0 for the angle C_{sp3}-C_{sp3}-H would depend on whether the central C_{sp3} atom had 1, 2 or 3 hydrogen's bonded to it. This level of complexity is of less use in a force field containing more atom types than the Carbon and Hydrogen found in an alkene force field as the non-hydrogen atom types connected to the central atom will have a much greater effect on the valence angles than in the alkene case where the other atoms must by definition be carbons. This will be refered to as the 'full with substitution factor' option for determining angle related parameters.

At the start of this project there was only one method being used to calculate the relevant parameters[1]. The relevant equations vary the angle depending on the number of hydrogen's involved in the angle bend interaction and are as follows:

$$\tfrac{1}{2}k_\theta(ijk) = \frac{d_1\left\{d_2 + d_3\left(\left|X(i) - X(j)\right| + \left|X(j) - X(k)\right|\right)\right\}}{l_0(ij)l_0(jk)}$$

$$½k_\theta(ijH) = \frac{d_4 \times d_1\left\{d_2 + d_3\left(|X(i) - X(j)| + |X(j) - X(H)|\right)\right\}}{l_0(ij)l_0(jH)}$$

$$½k_\theta(HjH) = \frac{d_5 \times d_1\left\{d_2 + d_3\left(|X(H) - X(j)| + |X(j) - X(H)|\right)\right\}}{l_0(Hj)l_0(jH)}$$

Where d_1-d_5 are parameters that can be optimised in the same way as c_1-c_4. X(i) is the electronegativity of the atom i. This will be refered to as the 'calculated' option for determining the valence angle bending force constants.

The formulae for the calculation of the strain free reference angles similarly depend on the number of hydrogen's and are as follows:

$$\theta_0(ijk) = \quad \theta_0(j)(max)$$

$$\theta_0(ijH) = e_1 \times \theta_0(j)(max)$$

$$\theta_0(HjH) = e_2 \times \theta_0(j)(max)$$

Where e_1 and e_2 are optimisable parameters and $\theta_0(j)(max)$ is the maximum bond angle for the central atom type j. This method will be refered to as the 'central times substitution factor' option for determining the strain free valence angles.

This method comes from the observation that most force fields show the pattern $\theta_0(ijk) > \theta_0(ijH) \approx \theta_0(Hjk) > \theta_0(HjH)$. When this was implemented it was found that is was fairly effective but was deficient in that it failed to take into effect the degree of substitution on the central atom.

It is obvious that in a situation involving a C_{sp3}-C_{sp3}-C_{sp3} reference angle that this angle would increase with the number of hydrogen's on the central atom and this is seen in the values in a force field[4] where the degree of hydrogenation is taken into account (No hydrogen's 109.5, 1 hydrogen 110.1 and 2 hydrogen's 110.4). The above system for calculating the reference angle will come up with only one value of θ_0 for all of these cases.

What was required was a system for calculating angles that took account of all the atoms connected to the central atom. The system that was implemented started not with values for individual angles but with a total for all the angles about the central atom. Each angle was then inspected and a weighting value calculated. The final reference angle for each angle is then as follows:

$$\theta_n = \text{Total Angle} \times \frac{\text{Weighting for angle n}}{\text{Total of all weightings}}$$

It was necessary to determine an empirical formula for the weighting to be given to

each bond angle. The equation for a bond angle i-j-k that was eventually used to calculate the weightings is as follows:

$$\text{weighting} = (S_i + S_k) / (l_0(ij)^a \times l_0(jk)^a)$$

Where S_i and S_k are size factors for atoms i and k, a is a general parameter of around 0.33. All of these can be optimised like normal parameters. This method will be refered to as the 'calculated' option for determining the strain free valence angle parameters.Table 4.1 shows a comparison of typical θ_0 values from all these methods of determination for a C_{sp3} atom connected to two other C_{sp3} Carbons and two Hydrogen's.

Table 4.1 Comparison of Strain free angle methods

	Angle		
Method	Csp3-Csp3-Csp3	H-Csp3-Csp3	H-Csp3-H
Central Only[2]	110.5°	110.5°	110.5°
All in Angle[3]	109.5°	----- *	----- *
Angle + Substitution[4]	110.4°	109.2°	109.1°
Angle x Sub Factor[1]	110.5°	108.3°	105.0°
From Total Angle	109.8°	109.6°	108.2°

* The Cosmic paper lists no angles containing Hydrogen's

4.4 Torsional Twist

The torsional term is very important in any force field and it is thus perhaps unfortunate that in a full force field it has by far the largest number of required parameters. For example, in the case of the MM2 force field[5] each possible set of four atoms involved in a twisting interaction can have up to three torsional barriers relating to different orders of rotation which are combined to give the overall torsional energy profile.

The very lowest level that has been used to determine the barrier height for a torsional rotation is that depending only on the bond multiplicity of the central bond[2]. So all single bonds would have one barrier height and all conjugated bonds another. This system obviously reduces the number of parameters to a minimum but is not usually considered practicable for producing accurate results.

The lowest reasonable level for setting the values of the torsional barriers is where the barrier height it determined by the atom types of the two central bonds and the bond multiplicity of the bond connecting them[3]. This method will be refered to as the 'central two' option for determining the torsional barrier parameters.

This method can be improved with a small increase in the number of parameters by using the technique of employing these general (or generic) parameters unless more specific ones are declared[1]. For example, in the case where the central bond is a C_{sp3}-C_{sp3} single bond the one fold rotational barrier V_1 would normally be set to zero. However, in the case where the full torsion angle is C_{sp3}-C_{sp3}-C_{sp3}-C_{sp3} better results are achieved if this V_1 parameter is given a small value and so a special value for V_1 will be specified in this situation[1].

The full situation involves specifying values for the barriers depending on all the atoms in the torsion angle. If this approach is used it will give rise to the greatest number of parameters in a force field and so is a situation that is to be avoided if at all possible. There are two variations on the full situation that are used in the optimisations. The first of these groups all the possible full torsion angle into distinct groups and uses a single value for each of these. This approach was used for the alkene force field where it was possible to use the groups defined in the WBFF[4]. This method will be refered to as the 'full (as WBFF)' option for determining the torsional barrier parameters. The second variation has seperate parameters for each of the possible combinations of 4 atoms in a torsion angle, except that a single barrier is used for all rotation about a C_{sp2}=C_{sp2} double bond. This method will be refered to as the 'full (trail)' method of determining the torsional barrier parameters.

Little work has been done on the calculation of the torsional barrier parameters. This is probably as the torsional barrier is a more abstract concept than say bond angle, and values vary more between different force field than other parameters. There is also the problem of finding the underlying patterns which any equations would have to reproduce.

4.5 Non-Bonded Interactions

No matter which formulae are used each possible non-bonded interaction between two atoms requires at least two parameters. If calculation is not to be used then there are two ways that these parameters may be specified.

The first of these comes from the fact that though there is a need for different atom types to cover both different hybridization and different situations within a molecule, when it comes to the non-bonded interactions it is reasonable to treat all the atom types of the same atomic type the same[3]. So, rather than have different parameter values for the interactions for the atom types C_{sp3}, C_{sp2} etc., only the parameters that relate to a general carbon atom would be required.

In the second case all the atom types are treated as different, i.e. the parameter values used for the C_{sp3}-C_{sp3} interaction will be different from those used for the

C_{sp2}-C_{sp2} interaction. These methods are divided up depending on the form of the equation used. The situation that will be refered to as 'A6-B9' uses a Lennard-Jones potential function as shown in 2.3.4.2 where the power term for the repulsive section is 9. Similarly 'A6-B12' uses a power of 12 at that point and 'A6-Exponential' uses a Buckingham function, as also shown in 2.3.4.2.

Calculation can be used to reduce the number of parameters in two similar ways.

In the first of these the parameters used are calculated from primary parameters associated with the atom types involved. Thus, for example, the minimum energy distance r_0 for any interaction can be calculated from the summation of the relevant radii of the two atom types involved. An example of this is the COSMIC[3] force field where the non-bonded equation used is a follows:

$$E_{VdW} = f\left[\frac{-2.25}{rr^6} + 8.28\times10^5 \cdot e^{\left(\frac{-rr}{0.0736}\right)}\right]$$

Where rr is the distance between the atoms divided by the sum of the van der Waals radii and f is the product of two 'well depth' parameters relation to the atom types involved, i.e.:

$$f_{CN} = f_C \times f_N$$

This method is refered to as the 'cosmic' option for determining the non-bonded parameters.

In the second case it is assumed that the various parameters for the interactions between different atom types can be found from the geometric mean of the parameters used for the interactions of the two atom types with themselves. For example, the parameters required for the C_{sp2}-H interaction could be calculated from the geometric mean of the parameters used in the C_{sp2}-C_{sp2} and H-H interactions[1]. An example of this approach used in the investigation detailed in chapter 7 is 'A6-B9 (Geometric Mean)' which uses a Lennard-jones function as given in 2.3.4.2 but uses the geometric means to determine the parameters of those A and B parameters for interactions between different atom types.

4.6 Out of Plane Bending

The out of plane bending force constant parameter is usually only determined by the type of the central atom and as there are few relevant atom types the number of parameters needed will be small. The main way to reduce the parameter numbers would be to use the same force constant for all cases.

4.7 References

1 D.N.J.White, J.N.Ruddock and P.R.Edgington, *Molecular Simulation*, 1989, **3**, 71.

2 E.K.Davies and N.W.Murrall, *Computers Chem.*, 1989, **13**, 149.

3 J.G.Vinter, A. Davis and M.R.Saunders, *J. Comput.-Aided Mol. Design*, 1987, 1, 31.

4 D.N.J.White and M.J.Bovill, *J. Chem. Soc. Perkins* **II**, 1977, 1610.

5 N.L.Allinger, *J. Am. Chem. Soc.*, 1977, **99**, 8127.

CHAPTER 5

OPTPARAM, an occam Program to Optimise Any General Molecular Mechanics Force Field

5.1 Introduction

It has been shown in the previous chapters that there is a large range of possible techniques for reducing the number of parameters needed to generate a force field. Until now there has been no way to compare the various combinations of these methods within a reasonable time period. To overcome this problem the program OPTPARAM was developed. The current version of OPTPARAM is written in the occam language and runs on a transparallel minisupercomputer consisting of 24 T800 transputers with a total of 96 Kbytes of memory.

The general approach behind the program is that a computerised general minimisation technique can be applied to optimise the correlation between the structural properties and energies produced from molecular mechanics minimisations with the same properties which have been determined experimentally.

In this case the minimisation technique used is a variation of that used within the molecular mechanics minimiser, namely the Newton-Raphson method. The only variation is that during the molecular mechanics minimisation the calculated shift is always implemented whereas in the parameter optimising routine the best situation is selected out of the calculated shift and also the points used in the calculation of the numerical derivatives.

The optimisation is done with respect to three different types of experimental data.

The first of these is the set of structural properties of lengths (mainly bonded but some non-bonded), bond angle and torsion angle for selected molecules.

Second is the energy difference between isomers. For example, the program will attempt to optimise the energy difference between gauche and anti butane to be within the range of the experimental value[1] of 0.628 ($\pm$ 0.047) Kcal Mol^{-1}.

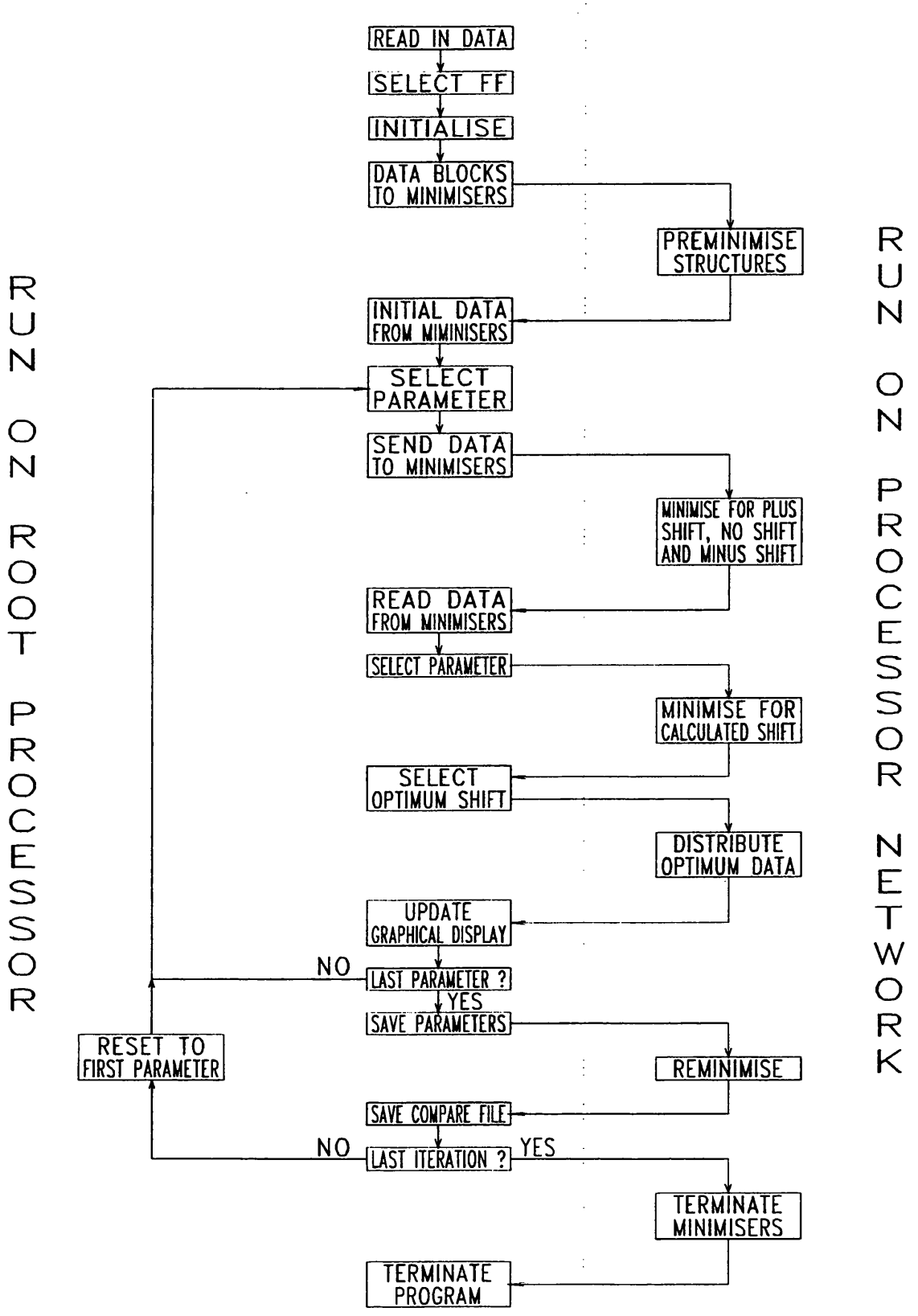
The third measure used in the optimisation is that of heats of formation. In this the assumption is made that various groups or bonds each contribute a fixed amount to the heat of formation. The current calculated energies can be used with a least square's technique to calculate the best current values for these fixed enthalpy increments. With these enthalpy increments and the steric energies it is then possible to calculate values for the various heats of formation. These predicted heats of formation can then be compared with the experimental values.

For each of these a residual factor (section 5.5.4) is calculated which gives a measure of the difference between the experimental and calculated properties, taking into account the size of the experimental error.

The residual that is minimised by the program is a combination of the above three measures. How these various measures are weighted to produce the final

measure is determined by the user at the beginning of a program run.

Figure 5.1 Program Structure.



5.2 Program Overview

The main structure of the program can be seen in figure 5.1.

Before proceeding to explain the operation of the program it is advantageous to introduce terms relating to the loops at different levels within the program.

The outermost loop is that during which each parameter is optimised in turn. As it covers all the parameters that make up the current force field, this will be called the force field loop.

The next operation is the optimisation of each parameter. It will be seen later that, in certain situations, an individual parameter optimisation may be repeated iteratively, but only when the results of each optimisation meet a criterion involving the fall of all residual factors. Usually though, each parameter is optimised only once during each force field loop and so this will be called simply a parameter optimisation and so shall not be considered to be a loop.

Within each parameter optimisation various shifts are made to the current parameter and for each of these a comparison of certain properties with experimental data is required. To do this the structures and energies which will result from the now altered force field must be found. This is done by minimising the force field generated energy of all the relevant molecules until one of several possible criteria is met. Thus the molecular mechanics minimisation routine iteratively loops for each molecule until it meets one of the criteria. The loop at this level will be referred to as the molecular loop.

OPTPARAM has so far been described as if it were a single program. In the currently implemented occam version though it consists of two separate programs. This division is required because of the way that occam is implemented, as it is difficult to boot one section of the code to the network and then start another section running on the root processor without keeping the two sections of code discrete.

OPTPARAM in this case is the name given to the main control section that runs on a transputer card within a PC and when activated replaces the TDS as the active program. The card used inside the PC is a Gemini GM8101 with 8 megabytes of memory and a 20Mhz T800 transputer. This program represents most of the development involved as it performs all the controlling, including deciding which parameters to minimise and which molecules to use, as well as handling all the user interaction, graphical display and file handling.

The second section is called MINIMSRS and consists of identical minimiser processes and a communication process and it is this section which is executed on each processor in the network. The transputer network at the end of the project consisted of 3 Gemini compute cards each of which contains eight 20Mhz T800

transputers, each transputer having 4 megabytes of memory.

The MINIMSRS section also includes the third main programming task which is a procedure called FFCONVRT. This takes the primary parameter values being optimised and expands them into the data required by the standard minimisers. These various data arrays, needed by the minimisers, will be referred to as the secondary parameters.

As there is only a fixed number of physical links possible between the transputers each program includes a communication task consisting of a multiplexer and a demultiplexer which decide which data to take and/or which to pass on to the next processor. For example, a copy of all initial data will need to be taken by each processor, whereas the data relating to a single molecule may be kept or passed on.

5.3 Program History

The program was originally written in FORTRAN 77 and ran on a Definicon 68020/68881 plug-in card in an IBM PC. This arrangement supplied only a fraction of the computing power required to run the program for the number of structures needed to produce a balanced force field. It was thought useful to determine whether the chosen approach would work, but it was only with the upgrading to the transputer network that it was possible to obtain any representative results in a reasonable time.

The language occam was used because, at the time that the transputers initially became available, it was the only language then available to run on a transputer system. So it was decided at the time, that, rather than wait for a FORTRAN compiler to become available, an attempt should be made to convert the original FORTRAN 77 program code to occam. This had a disadvantage in that the occam language, which is very different from FORTRAN, had first to be learnt and mastered. After this was done it proved to be beneficial as it is far easier to implement certain structures within occam than to use an adapted sequential language. This is especially true of a program like OPTPARAM, which uses a considerable number of ALT and PAR constructs, so that it would be necessary to produce and then configure together a large number of different tasks if it were to be implemented in FORTRAN.

Initially the program was written to optimise parameters using a specific form of the force field. As such this form was "hard wired" into the program in several locations. When it was decided that the program should be able to use different forms of the force field the first attempts to do this were made by making the changes to all these various locations in the program. This meant that when making a change to the

force field there was always a possibility that not all the correct changes would be made and that, because of the varied locations of the relevant code, even making a small change was a time consuming operation.

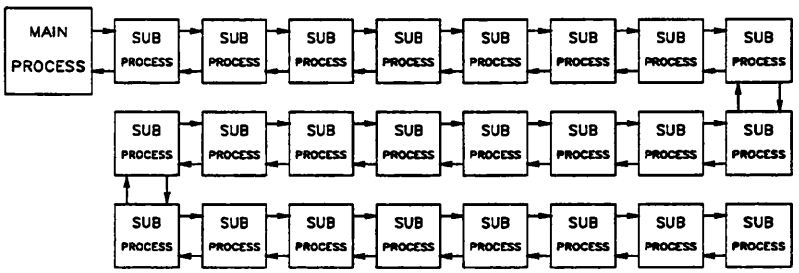
It was thus decided that, if possible, there should be a single procedure which would expand out the given primary parameters to produce the secondary parameter arrays required by the force field to calculate the energy. This has been implemented as FFCONVRT and has proven to be a rapid way of adding different methods and forms of equations for any of the terms used in the force field.

Besides these major changes there has been a continual upgrading of the program, with the regular addition of extra features plus detection and correction of coding errors. The result of this is that, as explained later in chapter 6, it is difficult to compare results obtained between different versions of the program, especially when different sets of data are used.

5.4 Network Configuration

As the minimisation of the molecules is a computationally intensive task, then, unless a commensurably large amount of data needs to be exchanged, the time spent processing will be far greater than that spent on communication. Because of this it was decided that the link configuration to be used would be a bi-directional pipeline (figure 5.2). Data being sent out from the control task is said to be going 'down' the pipeline using the 'out' channels. Similarly, data comes back 'up' the pipeline using the 'in' channels.

Figure 5.2 Pipeline Configuration of Transputers

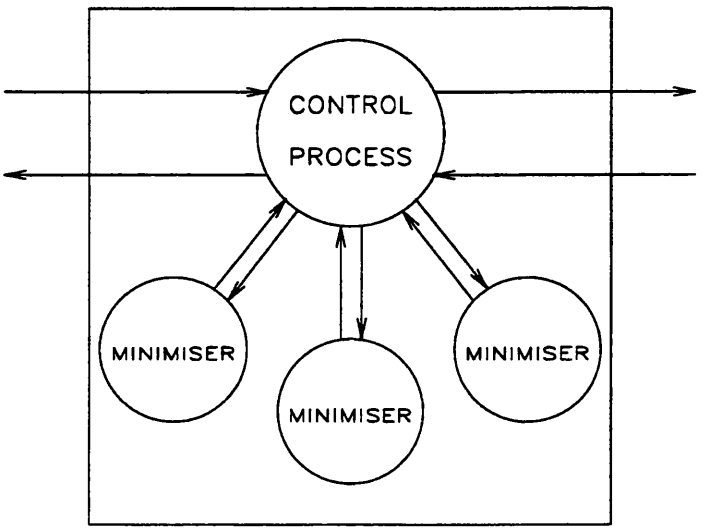


This configuration was chosen for the simplicity with which the message passing could be implemented. It is possible to reduce the mean number of links required to connect any two processes by using more complex configurations such as tree structures. In the current situation the communication time involved is not a significant contribution to the program duration and so it was decided that the extra complexity of using such a configuration would not result in a justifiable improvement in performance.

A simplified illustration of the configuration of the tasks on the worker

processors is shown in figure 5.3. It will be shown later that the communication task actually consists of two distinct sections, relating to the directions of data travel and that in certain circumstances' data is sent between them without passing via a minimising process.

Figure 5.3 Tasks on a Worker Processor



5.5 Program Methodology

Before going on to examine the occam implementation in detail, some of the more important points of the program methodology will be examined.

5.5.1 Minimisation Criteria

The Root Mean Square (RMS) value of the first derivatives for a molecule is effectively the average force at the current co-ordinate locations on its potential surface. So when the co-ordinates are close to their minimum energy configuration the RMS value of the first derivatives will also be small. Thus the RMS value can be used to give a measure of the deviation of a molecule from its nearest minimum energy orientation.

The initial idea for force field optimisation was that if the molecules could be constructed such that they represented the experimental data, then the RMS force could be used as a measure of the structural correlation, while the energies could be used in the other residual factors. Thus all that would be required would be to alter the parameters to optimise a combination of the RMS values and the energy residual factors. The advantage of this system would be its speed, in that, rather than doing multiple minimisation iterations to obtain the minimum energy configuration, it would simply require the calculation of the RMS value and the energy only once for each new parameter position.

This approach was rejected for two reasons. The main reason being that the structural data used to construct the molecules will always contain experimental errors. Thus, if it was possible to build a molecule to give the central value for each internal co-ordinate it would still be impossible to say that that was, in fact, the actual structure. For example, consider a bond angle that is given as 109.4°±0.6°, if the structure is built with the angle at 109.4° it is not possible to know if this is the correct angle, all that can be said is that it is the value in the centre of the possible range for the angle. In fact, with most molecules containing rings it proved to be impossible to build the ring such that all the internal co-ordinates are exactly at these central values. The other problem with this approach is that the sum of the first derivatives could be optimal if certain of the parameters went to zero as this would result in there being no energy for any configuration. This is overcome, to some extent, by using the energy residual factors, as these would increase if that were to occur.

The next concept was to produce and then optimise some residual factor between a minimised structure and the experimental data. This leads to two main questions, how to calculate the residual factor and how to decide when a molecule has been minimised enough with an adjusted force field for the comparison to be fair? The first question is covered in section 5.5.4 and only the second question will be considered here.

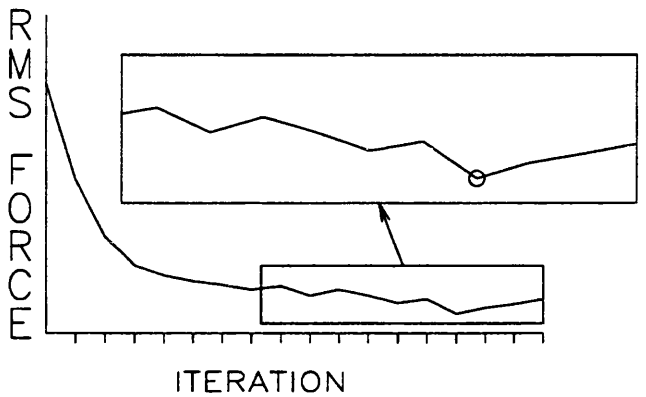
When the program was initially written it was obvious that certain parameters affected the molecules more than others and consequently it took the minimiser more iterations in these cases for the molecules used in the inverstigation to settle down to stable positions. If then, minimisation was to take place for an average number of iterations, there would be some cases where the molecule had reached the minimum energy structure a significant number of iterations before the final iteration and others where the structure would still be displaced from the minimum in the potential energy surface. So, if the minimisations were to take place for a fixed number of iterations, then that number would have to be fixed sufficiently high to make sure that all the molecules had reached their true minimum energy configurations.

As one of the main aims of the program is to optimise the force fields within the shortest reasonable time, this is obviously a major drawback. Thus it was decided that minimising iterations should take place until certain criteria had been reached, so that only those structures that needed it would be minimised for a large number of iterations.

When minimising to a criterion there needs to be some measure to which it is possible to compare the minimised structure to see if it has been minimised

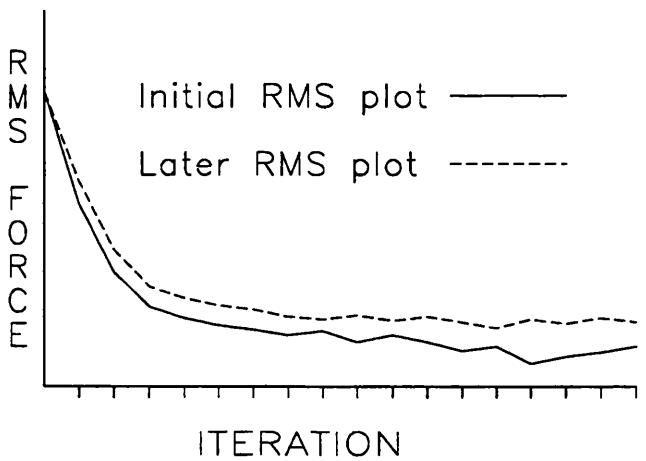
sufficiently. The obvious element to use in this case is the RMS force, because, as explained above, it gives a measure of the variation of the structure from the nearest energy minimum. Using the RMS force also has the advantage that it can be easily obtained from the minimisation and requires negligible extra calculation. The standard values of the RMS force used for comparison are the final values at the end of the initial molecular minimisation iterations as these are produced after a large number of iterations and so the minimum energy configuration should have been located for all the structures.

Figure 5.4 RMS Force against Iterations



The first attempts were made by trying to minimise until the new value for the RMS force was lower than the standard value. Observation of the RMS values at the end of each iteration (see figure 5.4) showed that, though at first the value comes down steadily, the value tended to vary in a random fashion about a final value. So, if the value of the RMS force at the end of the initial stage is one of the lower of these values then the minimiser may take an excessive number of iterations before the new RMS force can obtain this level, even if it has settled down at around the same level as that at the end of the initial minimisations.

Figure 5.5 Comparison of Different RMS Force Curves



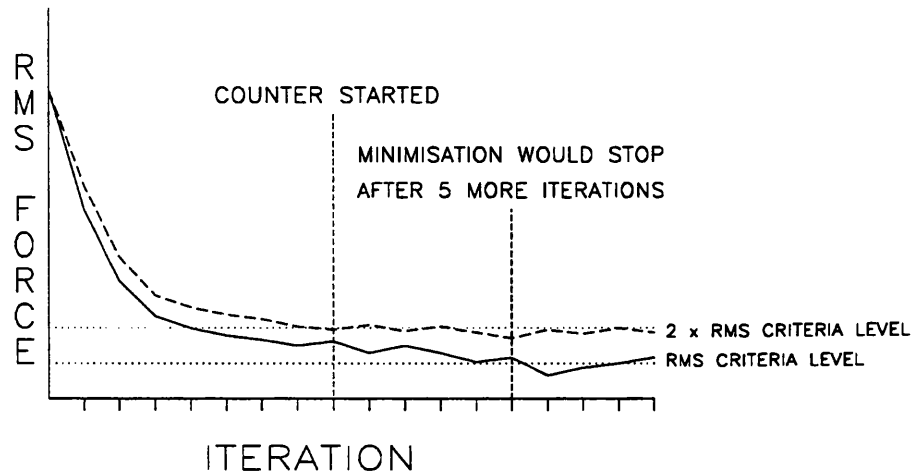
Also as the force field parameters change during the course of an optimisation, the value around which the RMS force will settle will also change. It is possible then, than what was a viable figure for the RMS criterion value at the beginning of the program, may be below that which is obtainable at a later stage in the optimisation (figure 5.5).

This problem has been overcome, to some extent, by having the program reminimise by the same number of iterations as used initially before each force field loop round the parameters. The RMS force values obtained from this are then used for the following force field loop.

There is still though a high possibility that there will be enough deviation in the RMS force level during the course of a single force field loop for this problem to cause the program to run excessively slowly or even hang altogether. To take account of this problem, extra termination criteria are required. The first of these is an attempt to cope with the random noise effect.

In this case, when the value of the RMS force drops below double the standard value, a counter is initialised and this is incremented every iteration thereafter. The minimisation now has two ways that it can terminate. The first is still that the RMS force should drop below the standard value from the initial minimisations but there is now another criterion, in that the minimisation will terminate when this counter has reached a pre-set number, even if the RMS force is still not lower than the standard RMS value. Currently the counter limit is set at 5 extra iterations, though this is purely a matter of personal judgement and can be easily adjusted. The operation of this criterion is shown graphically in figure 5.6.

Figure 5.6 Minimisation Termination Criteria



With this criterion there were a few cases where the optimisation would still hang due to the force field parameter changes as shown in figure 5.5, because though the change in RMS force level may be small for single optimisations, there is an

additive effect during the loop and so the level late in the force field loop can be significantly different. To counter this, a final criterion was added which is purely a check on the total number of iterations and this criterion will cause the minimisation to terminate when this counter has reached a pre-set value. This upper limit for the number of iterations is currently set at 400 iterations, but again this could be changed, if considered appropriate.

5.5.2 Load Balancing

When running a program that operates in parallel on many processors at once, it is obvious that to get the best out of such a system would require a situation where all the processors are running all the time. This would, ignoring communication overheads, give us the ideal situation of linear acceleration (i.e. a program running on twenty processors will be twenty times as fast as the same program running on a single processor). This may be nearly possible, for example, in situations where standard operations are being done on an array where the array can be divided up evenly between the processors. In the current situation though, there will be two major time consuming operations likely to be taking place. The first of these happens at the beginning of each force field loop and requires minimisation of all the molecules for a fixed number of iterations. For this situation it is possible to calculate a good estimate for the duration of the minimisation for each molecule by finding out in each case the number of interactions of each type.

The second time consuming situation is during the optimisation of the parameters, in which the molecules will be being minimised to certain criteria and in this case it is impossible to predict how long each minimisation will take. How then is it possible to implement load balancing for this situation?

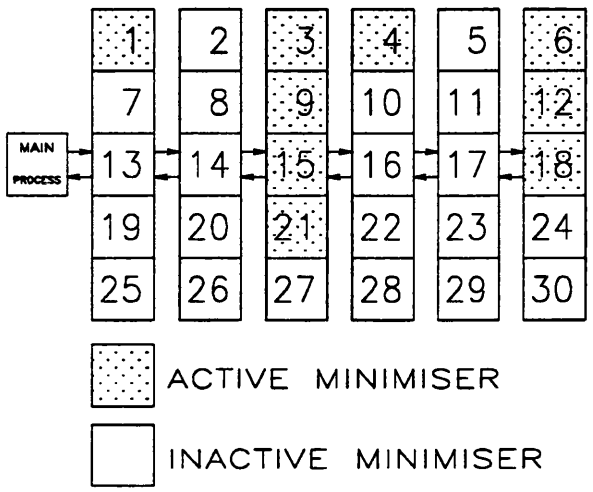
Initially the simple estimate for the time per iteration was used, because at the beginning it was not obvious how much of an effect the minimisation to a criterion, rather than for a fix number of iterations, would have. This allowed, at least a small degree of load balancing, while other approaches were tested. The system used at the time simply tried to get the summation of these estimates on each processor to be as close as possible. After this distribution of structures was calculated each structure was effectively fixed on a single processor.

Besides not knowing the duration of any minimisation the other main problem with this system is that it does not allow us to take full advantage of the fact that, for each parameter, only those structures affected by that parameter need to be used in the optimisation.

It is possible, either by initial analysis, or by analysis of the results being returned from the minimisers, to find out which molecules are not required for the

optimisation of any particular parameter (see section 5.5.3). To get the best performance from the system it is obvious that not only should the program minimise only the relevant structures for that parameter but also that these structures should be minimised over as many processors as possible. In the initial scheme it is quite possible to have, for example, 10 structures being minimised, where 4 are on one processor, 3 on another, three other processors with one structure each and the remaining processors doing no calculations at all (figure 5.7).

Figure 5.7 Bad Load Balancing



It was decided that a more flexible form of load balancing should be tried, as any likely increase in the communication overhead from a different method would be more than compensated for by an improvement in the load balancing.

In the currently implemented system the data for all the molecules is initially sent to all the processors. The structures are then processed to produce a list that ranks the structures in order of their likely minimisation time, to be used to determine the order in which parameters are optimised. At the beginning of the first force field loop this is done simply by arranging the molecules in order of decreasing size. Later, during each parameter optimisation, a timer notes how long each structure has taken and for the next optimisation of that particular parameter these timer values will be used instead of the size to determine the order in which the molecules are sent out.

Once the order of the structures has been decided they are sent out to the network to be minimised for the current force field. What is actually sent down the pipeline is, in fact, a number relating to the selected molecule and another to specify which processor it is to be run on. The structures are sent out so that the first molecule and thus, hopefully the one that will be slowest minimising, goes to the furthest processor in the pipeline. This is done to try and reduce the communication time as it allows the closer processors to finish and return their values before the

values from those structures further up the pipeline are returned.

If the number of structures is less than the number of processors then it is just a case of sending out all the structures and waiting for the MINIMSRS tasks to return the relevant data.

When there are more structures than processors then when data relating to the result of a minimisation is received back from a processor that has one or more free minimisers, then that processor is sent another structure to be acted upon. This is done until all the structures have been sent out and all their respective data has been received back (figure 5.8). Care has to be taken when using many molecules that a molecule is not sent to a processor on which all the minimisers are already being used. A check to stop this from happening is implemented by sending a flag back with the minimised data which informs the control program if that processor has any spare minimiser tasks remaining. If the flag shows that the processor is full then the control program will wait until the data arrives back from the first unfilled worker task before sending out the next data block. This approach is only needed for the optimisation loops, as the initial minimisations do not require any data to be preserved after returning the minimised data and so only a single minimiser section is required.

Figure 5.8 Good Load Balancing

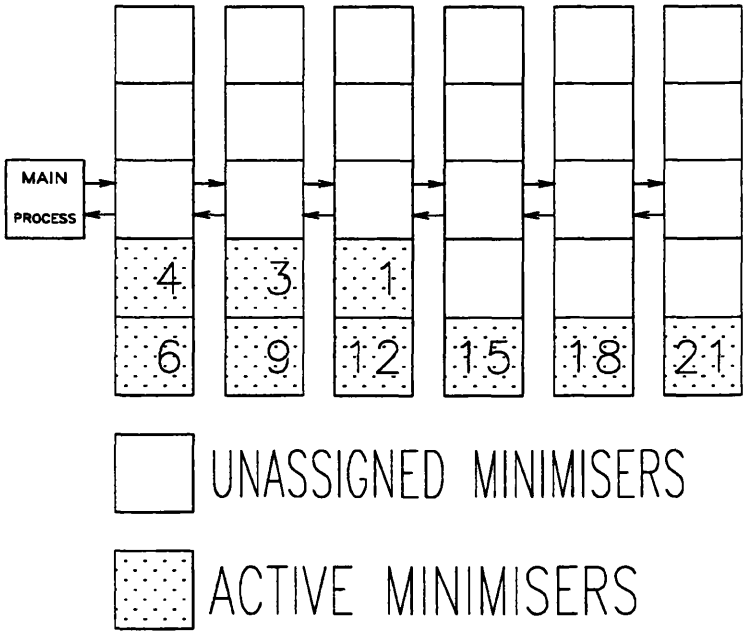
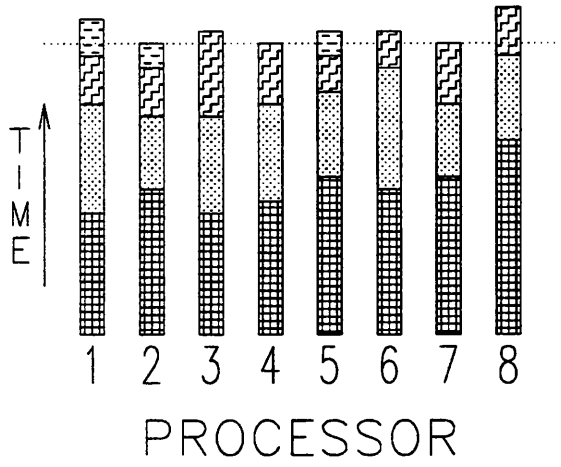


Figure 5.9 shows graphically, in a fairly idealised way, how sending the quicker minimising molecules last helps level off the overall load balancing, where the change in patterns represents a different molecule being minimised.

Figure 5.9 Hopeful Load Balancing Result



5.5.3 Determining Affected Structures

Due to the large number of variations in the force field used it is counter-productive to try to find out at the beginning which parameters affect which structures. This is because the whole system is based on the single procedure FFCONVRT which takes the current primary parameters and then uses them to produce the secondary parameters needed by the minimiser. To attempt to decide which primary parameters affect which structures would make things significantly more complicated, as a procedure would be needed to investigate the values and the produced arrays for each structure to see which of the initial parameters are used in each case.

Instead a system is employed where this information is obtained dynamically during the first force field loop.

What occurs is that during the first time a parameter is optimised all the possible structures are used. When the minimisers return their associated values for the energy and structure it can be seen, that where there is absolutely no difference between the sets of data, then that primary parameter has no effect on that molecule and so in all following iterations that particular molecule will not be required in the optimisation of that parameter.

It is also possible that the force field will contain parameters that are not used in any of the current structures. For example, if an optimisation is taking place on an alkene data set but using a parameter block that contains parameters for situations involving O_{sp2}, it is obvious that these parameters will not have any effect on any of the structures. The program thus also looks for cases where no structures are affected by a parameter and in a case of this type that parameter will cease to be optimised as it is obviously not optimisable with the current data set.

There are several reasons why this is useful. Firstly there is the time saving, as after the first iteration only those structures that are actually needed will be minimised. Secondly, it gives an indication of the validity of any given parameter. This is because a parameter is likely to be more accurate and thus more useful if it is utilised in a large number of the structures. An example of this would be when the option is used of defining both the atoms involved and the degree of substitution when deciding on the strain free bond angle. In this case it may be found that there are several angles that are only used in one or two structures and because of this small sample the resultant optimised angles could not be considered to be as accurate as angles that are optimised using more structures.

After the first iteration a file is saved which records the obtained information about which parameters affect which structures. This allows for a run to be stopped and then later restarted without the need for the first force field loop to minimise all the structures for all the parameters.

5.5.4 Residual Factors

As stated in the introduction, there are three residual factors that are combined to produce the overall residual factor that the program attempts to optimise. A residual factor is simply a value relating to how close the calculated data is to that determined experimentally taking into account the experimental error. The actual residual factors quoted will be the average of all the individual residual factors for that measure. All three residual factors use the same form of formulae to calculate the measure:

$$\text{Quoted Residual Factor} = \frac{\sum_j rf_j}{j}$$

Where each individual residual factor is given as follows:

$$rf_j = \left(\frac{\left(\left| Calc_j - Obj_j \right| - Err_j \right)}{Err_j} \right)^2$$

If $rf_j <$ zero then $rf_j = 0$

Where j is the total number of experimental properties.

5.5.4.1 Structural Residual Factor

Even though most parameters will not affect all the structures the value taken for the structural residual factor is always averaged over all the structural properties, i.e. j is always taken as the total number of structural measures over all the structures.

An average could be taken over only the affected structures in each case, but

this would make it difficult to see how the optimisation is proceeding as there will be no consistent value for the residual factor running throughout the force field loop. To make sure that all the structural measures are up to date, the values for the structural measure for those molecules which are not being used are set to the values relating to the optimum shift applied in the last optimisation in which that molecule was used. The same is done for the energy terms, so that the routines used to calculate the energy measures all use the most up to date values.

To minimise communication the values for the structural residual factors are calculated on the worker tasks and then sent to the control program when required. The procedure used to calculate the structural residual factor is examined in greater detail in section 5.6.6.

5.5.4.2 Isomeric Energy Difference Residual Factor

The isomeric energy difference residual factor is calculated in the control program using the values for the energies returned from the minimisers. The equation used in the calculation is exactly the same as that used in the structure residual factor case, except j now represents the number of pairs of isomeric structures for which an experimental energy is known.

5.5.4.3 Heat of Formation Residual Factor

The energy that is obtained from a molecular mechanics calculation is the steric energy, E_S, of the molecule in its current conformation. This energy can be considered as the energy that would be measured, if it were possible to obtain the molecule stationary at 0K. To compare this value with an experimentally determined value such as the heat of formation $\Delta H°_f$, the effect of temperature on the molecule needs to be taken into account ($\Delta H°_f$ is defined as the heat absorbed or released upon formation of the compound from the standard states of the elements composing it). As the molecules are considered in isolation, the $\Delta H°_f$ required is that of formation of the molecule in the gaseous state.

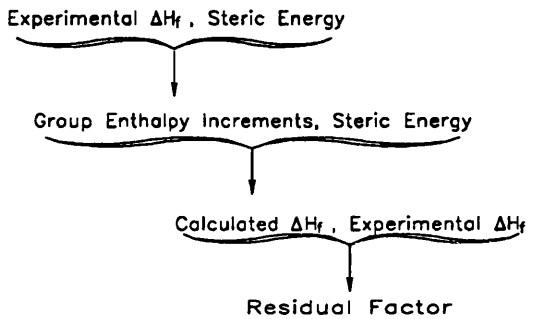
The effect of temperature on a molecule is to give it enthalpies of vibration, rotation and translation in addition to the vibrational zero point energy. It is possible to make these adjustments by making empirical corrections for specific groups or bonds.

This approach is called the group enthalpy increment scheme and assumes that a given molecular structure group or bond will make the same contribution to the heat of formation of a molecule regardless of the situation in the molecule in which that group or bond is found. Some justification[2] for this can be obtained by looking at the heats of formation of straight chain alkanes where the average change in the

ΔH_f for each CH_2 group going from C_2H_6 to $C_{12}H_{26}$ is -4.924 Kcal Mol^{-1}, with a standard deviation of only 0.19 Kcal Mol^{-1}.

The value for the heat of formation can thus be calculated by taking the steric energy of the molecule and adding to this the appropriate group enthalpy increments. The optimum values of these increments can be determined from the steric energies and experimental enthalpies by using a linear least-squares procedure.

Figure 5.10 Overview of Least Squares Operation



That is, that the experimental values for the heats of formation are used to calculate the optimum enthalpy increments for the current steric energies. These values are then used to produce the best values for the calculated heats of formation, which can then be compared with the experimental values to produce the residual factor (figure 5.10).

The calculation of the group enthalpy increments can be considered as solving a set of simultaneous equations where everything is known except the group enthalpy increments. For example, for 4 molecules that require various combinations of three enthalpy increments the equations are:

$$\Delta H°_{f1} = E_{s1} + I_1 n_{11} + I_2 n_{12} + I_3 n_{13}$$
$$\Delta H°_{f2} = E_{s2} + I_1 n_{21} + I_2 n_{22} + I_3 n_{23}$$
$$\Delta H°_{f3} = E_{s3} + I_1 n_{31} + I_2 n_{32} + I_3 n_{33}$$
$$\Delta H°_{f4} = E_{s4} + I_1 n_{41} + I_2 n_{42} + I_3 n_{43}$$

By subtracting the steric energy from the heat of formation and inverting the n matrix the least square values of the geometric increments can be obtained.

This is all done in OPTPARAM in the procedure "leastsq" using the Choleski process (see appendix A).

5.5.4.4 Enthalpy vs. Energy

In both of the above cases, care must be taken when attempting to equate energy and enthalpy. The two are related by the equation:

$$\Delta H = \Delta E + P\Delta V$$

Under standard conditions of temperature and pressure the errors involved in approximating PΔV to zero are usually small. In the case of the energy difference between isomers the difference is actually in the free energy but again, under normal conditions the term which relates the two can usually be ignored.

5.6 MINIMSRS

5.6.1 Task Configuration on Worker Processors

The minimisers used on the worker tasks are a version of the CHEMMIN molecular minimiser. Extra code has been added to allow for the receiving of the various data blocks from the communication task, to act on them accordingly and then return any results that may be required. The communication task consists of the multiplexing and demultiplexing processes that communicate with the nearest processors in the pipeline and with the minimisers on the current processor. A slight alteration has also been made to the minimiser sections to correct a problem relating to the use of numerical derivatives in the minimisation.

5.6.2 The Communication Task

All communication in OPTPARAM takes place as blocks of data that are passed around the network, most of which are sent from and to the main control program. It is the job of the communication tasks to take all these data blocks passing both up and down the pipeline and also to and fro from its associated minimisers and to make sure that they are all passed on or acted upon correctly.

With any block of data passing around the system the first byte is always a data tag telling the communication tasks how the following data in the block should be treated. In many cases this initial tag will be followed by more information that the communication task will need before it can process the data packet successfully, for example, the number of the processor on which the data is to be used. This is more true of outgoing data where a destination may have to be specified, than ingoing data where the destination will always be the control program.

So, for example, an outgoing instruction telling processor 6 to minimise structure 10 with shifts of 0.001 to parameter 35 will be sent down the pipeline as:

2	declared in datatags library as d.param.num
6	processor number
10	molecule number
35	parameter number
0.001	required shift in parameter

and a typical ingoing data block would be:

8 declared in datatags library as d.energy.dif
10 molecule number
1.500 structural residual factor
8.784 molecular energy

From the initial data tag the communication task can see that this data packet relates to the results of a minimisation and it is able by inspection of this tag to correctly route the complete packet to its proper destination at the control process.

When receiving a data packet the communication task initially looks at the first data value and using an IF construct it proceeds to execute the communication process relating to that particular data tag.

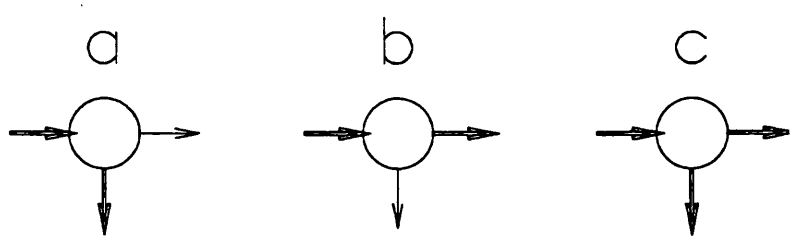
Some communication tasks are very simple, such as the one where all the initial data is sent out. In this case data is received and passed into the relevant arrays and if the current processor is not the last processor in the pipeline then the data is passed on down the pipeline to the next processor.

In the first example above however, another IF construct is required to see if the data is to be taken by the current processor or passed on down the pipeline. So, in this case, if the number of the current processor is less than 6 the data block will be passed on, but on reaching processor 6 it will instead be passed down to the first inactive minimiser.

In summary, for data coming down the pipeline there are 3 options for the communication task (figure 5.11).

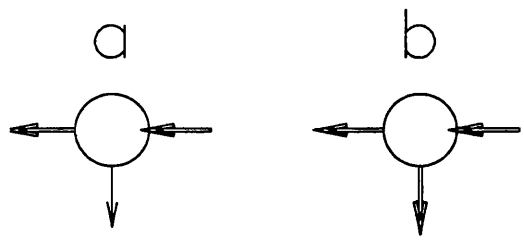
a) Pass on data down pipeline to required processor.
b) Pass data to a minimiser on this processor.
c) Store data in general arrays and pass it on.

Figure 5.11 Data Passing Down Pipeline



In the third case the process is receiving either the initial data or update data from other minimisers both of which need to be accessed from all the minimisers.

Figure 5.12 Data Passing Up Pipeline

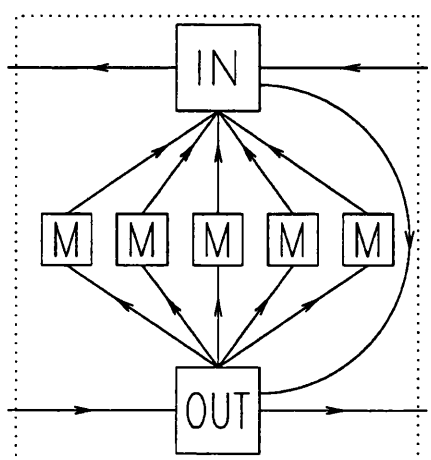


All data that is passed back up the pipeline is always passed on until it reaches the control program. The only variation is, as shown in figure 5.12, where in a few cases (b) the communication task takes a copy of the data as it passes through, as opposed to the normal case where it is just transferred on towards the control process using temporary variables (a).

Usually data is only sent in one direction by any one section of the communication task. The exceptions to this are the few situations where data needs to be updated on all the processors. This happens after the optimum shift has been decided and so the relevant co-ordinates must be transmitted from their current processors to all the other processors. This is a simple operation for the processors that are between the one with the original co-ordinates and the control processor as they need only take a copy of the data as it passed back up through the 'in' channel. The problem occurs with those processors further down the pipeline.

Rather than send all the data back to the control program and then have it transmitted back down the pipeline to processors past the originator, it was instead decided that the relevant data should be passed both up and down the pipeline in the same operation. This is done using a channel between the sections dealing with data going down the pipeline and data going up the pipeline. An extra copy of the data is sent through this channel from the up section to the down (figure 5.13).

Figure 5.13 Process Configuration on Worker Processor



This channel is part of an ALT construct with the normal 'out' channel and lets data pass into the data stream going down the pipeline as if it had come from the control program.

5.6.3 Variable Scope

The scope of the variables on each processor can be divided into two groups. The first are those that are declared at the very beginning and so are active for all of the minimisers and the communication task. This group contains:

The experimental geometric properties for all the molecules and their errors.

The current co-ordinates for all the molecules and their other properties such as atom types and charges, and bond connections and multiplicity's.

The current sums of the first derivatives which are used in the criteria for halting a minimisation.

The other group are those variables that are only declared in the minimisers, so that each minimiser on the processor will have its own distinct set of these variables. This set of variables contains:

The combined arrays ct and cct which after a force field conversion contain the current arrays of:

ct reference lengths, bond stretch constants, non-bonded parameters, torsion angle periodicity's.

cct conjugated reference lengths and conjugated bond stretch.

The arrays containing the information relating to which angle bend and torsional twists each atom is involved in and the parameters to be used for each of those angles.

Any out-of-plane bending information.

The history of the ct and cct arrays goes back to the original occam code for the minimiser when it was found that a system limit was reached when an attempt was made to pass all the separate arrays relating to force field parameters to a procedure. To overcome this problem these arrays were equivalenced to parts of two large arrays ct and cct that could then be passed to the procedure as the problem related to the number and not the size of the arrays. Originally the form of the force field was fixed to that used in the CHEMMIN minimiser and so these arrays were kept. When the program was altered to use any form of force field it became

impossible for sections of these arrays to be used as they had been previously envisaged. For example, it was originally possible to look in a section of the ct array to find what the barrier would be for a certain torsional twist, whereas, after the program was generalised, it was impossible to say how many such parameters there would be. So now such values are stored in the primary parameter list and are placed into the required array when the FFCONVRT program is run, rather than being passed through with the other information in the ct and cct arrays.

5.6.4 Program Operation

To describe the program operation it is easiest to examine each of the processes associated with the main IF statement of the communication task controlling data coming down the pipeline.

The first contains the control processes for processing data that is coming down the pipeline. This is the fold in which most of the operations take place and consists of the main IF construct. The second section relates to the previously stated case where data, such as co-ordinates, needs to be updated on processors further down the pipeline than the current processor. Thus this fold simply copies the required data coming in from the back.down.pipe channel and sends it to the 'out' channel.

The first fold contains the following code:

```
IF
  comm.type = d.initialise
    ... send initial data down pipeline
  comm.type = d.do.initial
    ... do initial preminimisation
  comm.type = d.new.set
    ... tell comm task that a new set is about to start
  comm.type = d.param.num
    ... send parameter to vary + shift
  comm.type = d.shift.calc
    ... send calculated shift
  comm.type = d.coord.type
    ... send final shift type
  comm.type = d.update.coords
    ... update co-ords from elsewhere in network
  comm.type = d.remin
    ... do reminimisation
  comm.type = d.update.parameter
    ... update parameter
  comm.type = d.finish
    ... finish all
```

Looking at each section in turn:

5.6.4.1 d.initialise

Due to the way that the load balancing works (section 5.5.2) each minimiser is not assigned a specific molecule, but rather any minimiser process should be able to be applied to any of the structures.

This means that each processor must contain all the relevant data a minimiser will need to work on any structure. At the beginning of any run of the program all this data must be sent down the pipeline to each processor and this is accomplished by the task relating to the d.initialise data tag.

A large data packet will be sent down the pipeline from the control program. Its first byte will be the d.initialise tag followed by all the required data. Most of this data is stored in the general arrays at the uppermost level. The only information that is passed down to the processors is the periodicity's of the torsional interactions, the electronegativities of the atom types and the initial parameter values. The reason for doing this for the first two is that the minimiser expects this data to be part of the general non-conjugated ct array and this section of the minimiser was not adapted during the modifications. The parameters are sent because they must be stored in an array at the minimiser level to avoid different minimisers on the same processor accessing a shared array that may be changing.

As well as the information mentioned above the data that is transferred can be divided into two parts relating to the parameters and to the structures.

For the parameters the following data is sent: number of parameters, parameter types, parameter identifiers, force field configuration, a flag to show if the off-diagonal terms in the non-bonded interaction are to be geometrically averaged from the diagonal terms and the number of initial (and reminimisation) iterations.

The following data is transferred for all the structures: number of atoms, initial co-ordinates, atomic charges, atom types, atom connectivities, current sum of first derivatives, bond multiplicity's and all information relating to experimental geometric properties.

All this data is stored in the appropriate arrays and if the current processor is not the last processor in the chain the data packet will be passed on to the next processor in the pipeline. As this process does not produce any results there will be no related data package to be returned to the control program.

5.6.4.2 d.do.initial

This data tag tells the communication task that initial minimisation needs to be carried out on a specified molecule. The tag is followed by the processor number on which the minimisation is to take place and the structure that is to be minimised on

that processor.

This data packet will pass down the pipeline until it reaches its assigned processor. At this point the data tag and the molecule number will be sent to a minimiser for the initial minimisation to take place.

The tasks involved in minimising a molecule, either for a fixed number of iterations as in this case or until a criterion is met, are covered later in this chapter and will not be discussed here. When the minimisation has taken place, the following are extracted: a new set of co-ordinates, a new value for the RMS force (which are taken from the last iteration) and the final energy. The next step is to produce a value for the residual factor between the minimised structure and that determined experimentally. This is done using the compare procedure (section 5.6.6).

Once the comparison has been done the data is then ready to be distributed. The minimiser returns a data packet to the communication task consisting of a tag (d.initial.energy.dif), the structural difference measure, the energy, the RMS value and the new co-ordinates. As most of this data is required on all the processors the communication task sends this data packet both ways along the pipeline.

5.6.4.3 d.new.set

During the course of the minimisations used in parameter optimisations the communication task keeps a count of the number of structures on its processor. Before any new process of this sort can take place this number must be reset to zero ready for a new set of structures to be sent to the processor. The "d.new.set" data tag causes this to happen and, as with the initial data, it will be passed along the pipeline until it reaches the final processor.

5.6.4.4 d.param.num

To calculate the optimum shift for a parameter it is necessary to know how the various residual factors vary with small shifts to that parameter. To do this all the relevant structures need to be minimised for the central and shifted position of this parameter. This is done by sending out a data packet with this data tag for each molecule needed. The data following the tag inform the communication task on what processor the minimisations are to take place, the structure to be minimised, the parameter to be shifted and the parameter shift. When this data arrives at its required processor it is assigned to the first free minimiser. The minimiser does the minimisations for the three cases (no shift, plus shift and minus shift) and for each case it calculates the structural residual factor and the energy. Once all this has been done this data is sent back to the control program as the tag "d.energies.difs" followed by the molecule number, the processor number, a flag to show if that

processor is full or not, the three residual factors and energies and a flag that shows if that structure was affected by that parameter. The state of this last flag is decided by looking at the structural difference and energy values. If they each have the same value for all three different situations then that molecule is not affected by that parameter and so in later iterations there will be no need to minimise that molecule for that parameter. This data packet will be sent to the control process without needing to be copied by the intervening processes.

5.6.4.5 d.shift.calc

When the control program has received all the information relating to the shifts it will use the Newton-Raphson method to calculate a new shift that should produce a lower total value for the combination of the various residual factors. This shift is sent out down the pipeline and is passed on to all the active minimisers on each processor. At each minimiser the shift is applied to the relevant parameter and the current molecule is again minimised and the structural residual factor and energy are once more determined. These values are then sent back with the tag "d.energy.dif" along with the molecule number.

5.6.4.6 d.coord.type

After the control program has received the information relating to the calculated shift it can compare its total of the residual factors with that calculated for the three situations used in the calculation of the shift. The control program selects which of these situations produces the lowest total residual factor and a tag showing which situation this is, is sent down the pipeline following the d.coord.type data tag. All the minimisers store the co-ordinates for all the four cases (centre, plus shift, minus shift and calculated shift) and when the tag is passed to the minimisers the relevant co-ordinates are sent out to the network as well as being updated on the processor itself.

This again is a case where data is passed both ways along the pipeline so that the co-ordinates on all processors are updated. The data tag used is "d.coords" and this is followed by the molecule number and its co-ordinates.

5.6.4.7 d.update.coords

When the data packet starting with the data tag d.coords arrives at the communication task it is sent both ways along the pipeline. However, to distinguish between the two packets the data tag on the packet sent down the pipeline is changed to d.update.coords. On receiving this packet a processor will update the relevant co-ordinates and pass on the data to the next processor until it reaches the last processor.

5.6.4.8 d.remin

At the end of a full force field loop the control program may request that all the structures be reminimised for the current force field in a fashion similar to the initial minimisations. Again this will result in a data packet with the tag "d.coords" being sent out from the minimisers to be transmitted to all the other pipeline processors as well as the control program.

5.6.4.9 d.update.parameter

After the optimum shift has been sent to all the processors and the relevant co-ordinates have been distributed there is then the problem that the value for the relevant parameter must be updated on all the minimisers on all the processors. This is done using a data packet with the tag d.update.parameter, followed by the parameter number and its new value. The communication task passes this data to all the minimisers to update their parameter arrays and transmits it to the next processor until the last processor is reached.

5.6.4.10 d.finish

When an occam program finishes it is considered good practice to terminate all the processes that are running. In this case the program should terminate after the requested number of force field loops have taken place. To terminate the network processes the d.finish data tag is sent down the pipeline that sets the logical variable used in the WHILE statement to FALSE causing the communication task to terminate. Before terminating the communication task sends the data tag to all the minimisers so that they to terminate.

5.6.5 Minimiser Task

When a minimisation is requested by the OPTPARAM program what happens is as follows:

1) Any variations in the primary parameter list are made i.e. shifts in the parameter concerned.
2) A call is made to the FFCONVRT routine which converts the primary parameter list into the required arrays for the standard minimiser.
3) A call is made to the standard minimiser which will terminate after either a set number of iterations or when one of several possible criteria is met, depending upon the situation.
4) Any calculations on the final structure are done and any information (if required) is returned to the OPTPARAM program via the communication

task.

5.6.6 The Structure Residual Factor

To reduce the required communication each minimiser calculates the structure residual factor for its current molecule. The equation used to calculate the measure has been given previously in section 5.5.4.

The following is an extract from the code used to calculate the residual factor, showing in detail the fold relating to the bond lengths. The code for the other structural properties is similar, with the exception that the torsion angle case must take into account that -180° = 180°.

```
#USE clcprocs    1
PROC compare(VAL [mxgeop][4]INT gpropa,
             VAL [mxgeop]REAL32 gpropv,gprope,
             VAL [mxgeop]INT gpropt,
             [mxgeop]REAL32 gpropr,
             VAL INT ngeopt,REAL32 dif2,
             VAL [mxat][3] xo)    2
  VAL length IS 1 :
  VAL angle IS 2:
  VAL torsion IS 3 :
  REAL32 retval,dif :
  SEQ
    dif2 := 0.0(REAL32)    3
    SEQ i = 0 FOR ngeopt     4
      SEQ
        IF
          {{{ length
          gpropt[i] = length    5
            SEQ
              calc.dist(gpropr[i],xo[gpropa[i][0]], 6
                xo[gpropa[i][1]])
              dif := ABS((gpropr[i] - gpropv[i]))    7
              dif := dif - gprope[i]    8
              IF
                dif < 0.0(REAL32)
                  dif := 0.0(REAL32)    9
                TRUE
                  SKIP
              dif := dif /gprope[i]    10
              dif := dif * dif
          }}}
          ...  angle
          ...  torsion
        dif2 := dif2 + dif    11
:
```

The procedure needs to call calculation routines such as calc.dist (6) which are in a library called clcprocs. This library must therefore be USEd (1) at the beginning. When the procedure is called it has passed to it (2) the following data:

gpropv	containing the geometric property values.
gprope	the experimental errors for those values.
gpropa	contains the atom numbers involved.
gpropt	contains the type of each geometric property.
gpropr	will contain the results of the calculation of each property.
ngeopt	is the number of geometric properties.
xo	holds the co-ordinates.
dif2	will be the final measure of closeness value.

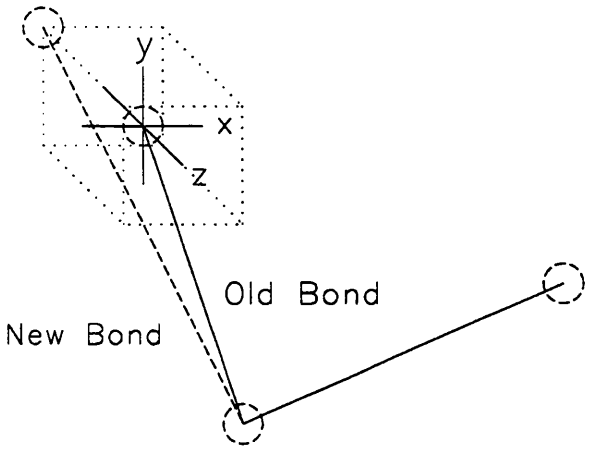
Initially the residual factor is set to zero (3). The program then loops round all the geometric properties (4). The next section depends on what the current geometric property is (5) but the approach is similar in all cases. Firstly the current value gpropr[i] is calculated (6). The difference between this and the experimental value gpropv[i] is then calculated (7). The function, as previously stated, is designed to measure how far from the experimental error range the current value is, so the dif value has the relevant error subtracted from it (8) and then if this results in a value of less than zero (i.e. within the error bounds) the value is set to zero (9). If the geometric property is outside the error bounds and so produces a non-zero value then this value is scaled by dividing it by the experimental error (10). This means that in a case where there is a large experimental error a large difference from the experimental value will produce a lower residual factor than a similar case where the experimental error is smaller. The value for this geometric property's contribution is then added to the running total (11).

5.6.7 Alterations to the Minimiser

The minimiser section used in MINIMSRS is basically the CHEMMIN minimiser[3] that was converted from FORTRAN by Noel Ruddock in our laboratory. One alteration has been made to cope with a problem resulting from the use of numerical derivatives.

This problem can arise when an initial bond length is substantially different from the result likely for the final minimised value. As the bond stretch force constants are generally large this will give rise to a large gradient. In these cases it is possible that the atom position will be moved to either extend or compress the bond as much as possible without other force values such as bond angle bending having much effect (figure 5.14).

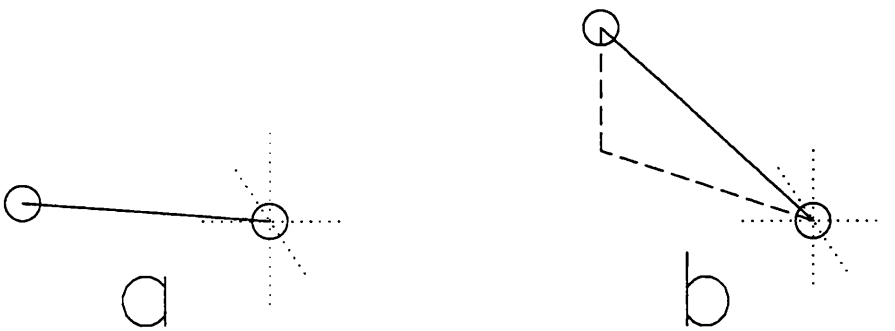
Figure 5.14 Maximum Shift Due to Large Bond Distortion



The result of this is that, when the bond comes to equilibrium, the bond angles it is involved in can be highly deformed. Whether or not the molecule can then recover from this deformation depends on the bond's direction (figure 5.15).

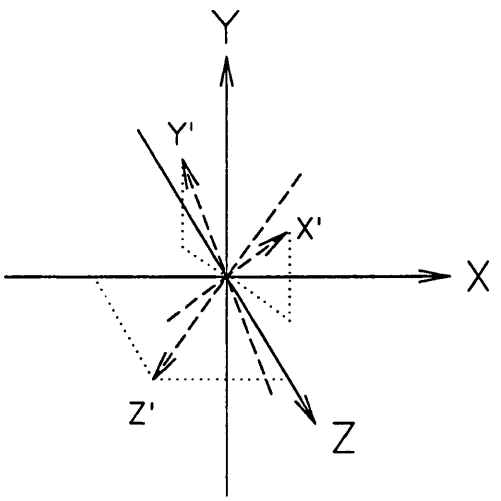
If one of the planes formed by the axis is roughly perpendicular to the bond (situation a, in figure 5.15) then during minimisation a gradient will be set up which involves little change in the bond length and the atom will move round towards its proper minimum energy position. The problem arises if none of the planes line up in this manner (situation b). Here all the positions used to calculate the gradients result in either extension or compression of the bond and the extra energy from these shifts overpowers those from the other contributions such as angle bending. The result is that the current position is seen by the minimiser to be near the bottom of a potential well and thus it gives the atom only a small shift, effectively fixing the atom in its current position.

Figure 5.15 Different Bond Directions



To over come this problem the axes along which the atoms are moved are skewed on alternate interactions (figure 5.16). The result being that an atom is unlikely to be in a false minimum for both sets of axes, thus allowing it to move towards its proper minimum position in all situations.

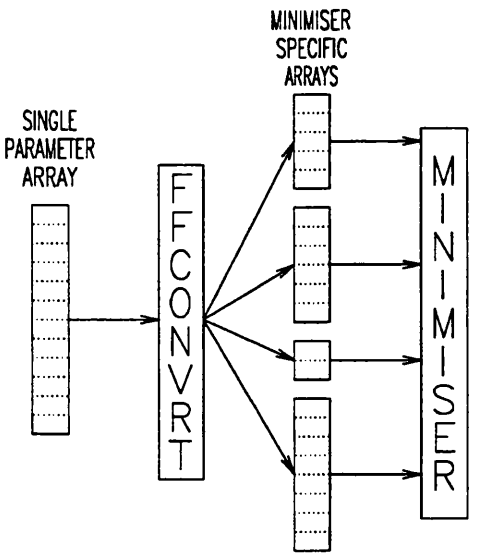
Figure 5.16 Skewed and Normal Axis



5.7 FFCONVRT

The FFCONVRT procedure takes a specific molecule, the current list of primary parameters and any control data from OPTPARAM and uses this data to construct all the arrays that a minimiser needs for minimisation to take place (figure 5.17).

Figure 5.17 Use of FFCONVRT Procedure



When the primary parameters are loaded from files each parameter has associated with it up to 5 values. Of these 5 values the first indicates the type of the parameter and the other 4 relate to the situation where that parameter should be used.

The parameter type can either be specific to the selected form of the interaction or general for the different forms of that interaction type. For example, there are parameter types relating to the calculation of bond angle parameters that are specific to

that way of producing the force field but there are also two non-bonded parameter types that are used for all the different forms of the non-bonded interactions.
The full list of the parameter types are:

0	Bond lengths
1	Bond stretch force constants
2	Non-bonded 1
3	Non-bonded 2
4	k_θ
5	k_θ'
6	Angle values
7	Torsional barriers
8	Out of plane bending
9	Bond force constant from length parameters
10	Atom weights for calculation of angle
11	a power term used in calculation of angle
12	Total angles for calculation of angles
13	Conjugated bond lengths
14	Conjugated bond stretching constants
15	Conjugated torsional barriers
16	Simple bond angle substitution factors
17	Bending parameters (k_θ) from electronegativity parameters
18	Parameters to adjust torsional barriers using electronegativities
19	Simple calculation of angle θ_0 parameters
20	Simple calculation of angle k_θ parameters
21	Lone pair parameters
22	One fold torsional barriers
23	One fold torsional barriers (conjugated)

5.7.1 Procedure Operation

The FFCONVRT procedure goes through each of the interaction types in turn expand out the selected primary parameters according to the force field form selected.

For the bond lengths the procedure is simply a case of scanning through the parameter types for the primary parameters and filling in the appropriate locations in the reference length (refine) array. i.e:

```
SEQ
  SEQ i = 0 FOR maxtyp
    SEQ j = 0 FOR maxtyp
      reflen[i][j] := zero32      1
  SEQ i = 0 FOR num.of.params
    IF
      what.param.is[i][0] = bond.length      2
        SEQ
          reflen[what.param.is[i][1]]
             [what.param.is[i][2]] :=
             param.values[i]    3
          reflen[what.param.is[i][2]]
             [what.param.is[i][1]] :=
             param.values[i]
      TRUE
        SKIP
```

1) All the reflen array elements are initially set to zero. This is so that, if an insufficient data set is accidentally used, then when the minimiser inspects the bond length array for a parameter which is not present it will find a zero and cause an error, stopping the program.

2) "What.param.is" is the array containing the 5 values about the parameters and the first of these contains the parameter type. So in this case only those parameters relating to bond length are used.

3) The next two elements in the what.param.is array show which two atom types relate to that bond and so these are used to index the correct locations in the reflen array. Both sides of the array are filled in this manner.

For the bond stretch parameters the program is currently set up to handle two situations:

```
IF
  wchcfg[bond.stretch] := full
    ... fill stretch constant array as per length
  wchcfg[bond.stretch] := calculated
    ... calculate entries into stretch constant array
  TRUE
    STOP    -- must be an error if neither selected
```

The first is that the values should be declared explicitly in full, in which case a process similar to the one above for bond length will be executed. In this case the type that will be looked for is that for bond stretch and it is the strpar array that will be filled rather than the reflen array.

The other situation involves the calculation of the bond stretch parameter using the formulae as described in chapter 4 and the required code for this is in the

second fold. The code within that fold is as follows:

```
{{{ calculate entries into stretch constant array
[5]REAL32 strpar :      1
SEQ
  {{{  get strength from length parameters
  SEQ i = 0 FOR num.of.params
    IF
      what.param.is[i][0] = stren.from.len
        strpar[what.param.is[i][1]] :=
           param.values[i]    2
      TRUE
        SKIP
  }}}
  SEQ i = 0 FOR maxtyp
    VAL rei IS reflen[i] :     3
    SEQ j = 0 FOR maxtyp
      VAL deq IS rei[j] :
      IF
        {{{ if zero then skip
        deq = zero32      4
          SKIP
        }}}
        {{{ if j smaller than i then skip
        j < i      5
          SKIP
        }}}
        {{{ if not calc stretch constant
        TRUE
          REAL32 skij :
          SEQ
            VAL deq2 IS deq * deq :
            VAL deqm1 IS deq - strpar[0] :
            VAL deqm12 IS deqm1 * deqm1 :
            IF
              (i > Hh) AND (j > Hh)     6
                skij := ((strpar[1] / deq2) +
                   (strpar[2] / deqm12) -
                   (strpar[3] / deq)
              TRUE
                skij := strpar[4] / deq2
             strcon[i][j] := skij
             strcon[j][i] := skij
         }}}
```

Where the following notes apply:

1) The array strpar is declared and dimensioned at the top to contain the required calculation parameters.

2) The strpar array is loaded with the appropriate values from the primary parameter array so that there is no need to search all of this primary parameter array.

3) The variable rei is assigned to a row of the reflen array. This means

that later on the selection of the bond length requires reference to a single rather than a 2D array.

4&5) There is no need to do the calculation if the bond length is zero, i.e. that particular bond length is not currently being used, or when j is less than i, as it is quicker to fill both sides of the matrix at the same time as the final matrix will be symmetrical.

6) The calculation varies depending on whether one of the atoms is a hydrogen or not, so it is necessary to test for this situation for both of the atom types involved in the selected bond length.

Similar set-up procedures are carried out for the conjugated length and strength parameters as well as for the non-bonded parameters with some parameters being just copied into the appropriate arrays, whilst in other cases some calculation may be required before the arrays are completed.

All these arrays can be filled without any knowledge of the molecule to be minimised. This is not true for all the other parameter types. In these cases the minimiser expects all the cases where interactions take place to be in a list such that it can quickly find out which interactions an atom is involved in and what the relevant parameters are. These arrays therefore need to be re-constructed each time that the FFCONVRT procedure is called.

This is done in two parts. The first is the scanning of the molecule to find all the possible valence and torsion angles, the second to either look up or calculate the relevant parameters for these angles.

It is of course possible to have a routine to find these lists and then just leave the FFCONVRT procedure to go through them updating the new parameters. As the time taken on the transputer to find these lists is small it was decided to make the FFCONVRT procedure recalculate all the various arrays. This means that all that is required at any point to make sure that the force field is up to date is one call to the FFCONVRT procedure. The advantage of recalculating all the lists is that it saves space on the processors as each processor would otherwise have to store the complete list of the valence and torsion angles interactions for all the molecules.

A full listing of the FFCONVRT procedure can be found in appendix A.

It should be noted that some of the code quoted in this and other sections has been slightly simplified to improve readability. The main difference is that to save space within the program many of the arrays are stored as INT16 variables and when these are used as indices in arrays they must be converted to the full INT. For example, what is given above as:

```
strpar[what.param.is[i][1]] :=
  param.values[i]
```

In the actual code this is given as:

```
strpar[(INT what.param.is[i][1])] :=
  param.values[i]
```

5.8 OPTPARAM

This section will initially examine the way that the program runs as seen from the users perspective and then go on to examine in more detail how the program actually functions.

5.8.1 Force Field Initialisation

Once MINIMSRS has been booted to the network and OPTPARAM started on the root processor the first stage is the initialisation stage where the user of the program can decide upon the form of the force field to be used, the data from which to start the optimisation and the conditions under which the run will take place.

After the initial information screen has been displayed the user is prompted for the form of the force field. The force field is divided into sections roughly related to the types of parameters used in the interactions. The sections used are:

1) Bond lengths
2) Bond stretch force constants
3) Strain free valence angles
4) Angle bend force constants
5) Non-bonded parameters
6) Torsional barriers
7) Out of plane bending parameters (and any other terms if used)

By use of the up and down cursor keys it is possible to see the various options, if any, for each type and then select the required method for that section. The program reads a file relating to the default force field which determines which method is initially displayed for each section.

Before going on to the next section it is also possible to fix the selected parameters, either for the complete duration of the run or for a certain number of iterations. This is useful if attempting to optimise different forms of one of the parameter types, for example, the form of the non-bonded interaction, as it is

possible to fix all the other parameter types for a few iterations allowing the unfixed section to come nearer to equilibrium before allowing the rest of the parameters to be optimised. This not only optimises the force field quicker but also reduces any unneeded movement of the other parameters as the changed section comes closer to equilibrium. The fixing is done using the '+' key. One press will fix the block for the whole iteration but any further presses after this will convert it to being fixed for a limited duration related to the number of presses. For example, pressing the '+' key 6 times will mean that the block will be fixed for the first 5 force field loops. It is possible to cancel the fixing option by pressing the '-' key before going on to the next section.

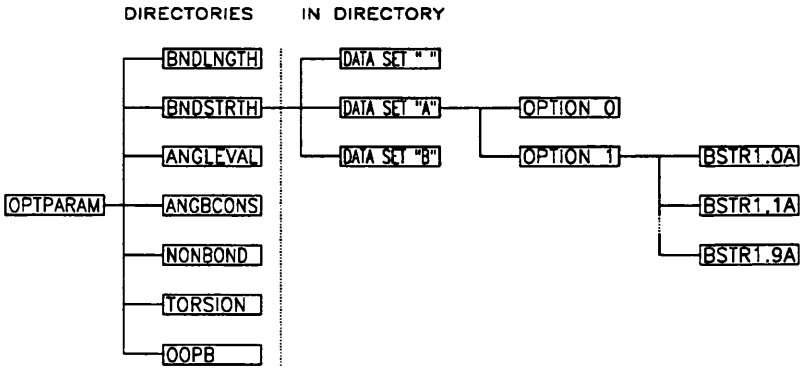
Once the form of the force field has been selected it is saved as the new default force field ready for the next run.

The program now prompts for the data set to be used. To explain what the data set relates to it is best to describe how the various files are arranged.

There are seven main subdirectories each relating to one of the sections above. For example, bond length parameter data will be in a different directory to that containing torsional barrier data.

These sections can then be divided depending on the method used for the interaction, for example, A6-B12 as opposed to A6-B9 non-bonded interactions. These divisions are implemented by using different file extensions. Each filename must also contain information relating to when the data contained was produced. To do this the files are numbered depending on the order in which the data was produced, i.e. if it is the initial data it will be labelled 0, if the first set optimised, it will be 1. This number is appended to the main part of the file name. It should be noted that the number relates to the next free spot available at the time and that the data was not necessarily optimised from the previous data set. This organisation is shown graphically in figure 5.18.

Figure 5.18 Relationship Between Data set and Filename



Data sets

The program uses a single character to determine which sets of parameters, called here data sets, are to be used.

The use of data sets has several purposes. The first and simplest is in time saving, as by dividing the data up into different data sets it means that the program will spend less time searching the disk. A new data set is usually started when adding a new atom type as it also helps to keep incompatible data blocks apart.

As stated the data set is defined by a single character, which is appended to the extension of the data block's filename. Thus a data block for an alkene force field might be BLEN30.0A whereas with the addition of data for O_{sp2} a similar file might be BLEN30.0B. So if an optimisation for molecules containing oxygen's is required the user can select data set B which will then avoid loading files which do not contain the required parameters.

In the above case 'BLEN' refers to the section, in this case bond length. The '0' in the extension shows which type of parameter is contained in the block and the '30' shows it is the 31st block (starting at zero) that has been written for that type. The 'A' relates to the data set.

Now that the methods for each section have been selected as well as the required data set the program searches the appropriate directories of the disc for all suitable data files. When these have been found the program takes the last file in each set and displays the accompanying information file (appendix C). If this is the required file it can be selected simply by pressing <return>, but it is possible to choose any of the data files by using the left and right cursor keys to scan through them. In each case the information file will be displayed to help the user decide which block to use. Once all the required files have been selected their contents are read and simultaneously displayed on the screen with a short pause between files.

The program then reads the files containing the rest of the data required for the optimisation.

Firstly there is the experimental data that comes in two parts relating to data particular to each molecule and then data relating to pairs of molecules. STRUCTUR.LST is a file read in by the program that contains a list of the files relating to the current molecules being used in the optimisation. The STRUCTUR.LST file also contains any data relating to the experimental energy difference between isomers within this group of structures. The isomers are identified by their locations within the STRUCTUR.LST file. An example of a typical

STRUCTUR.LST file is given in appendix D.

The files that are listed in the STRUCTUR.LST file contain the molecule's name and initial structure as well as the experimental data such as bond lengths and angles and the heat of formation (see appendix E for an example of the format). These files are stored in the MOLDATA subdirectory.

After this final data initialisation stage comes the loading of the constants that are not optimised. These consist of the electronegativities of the atom types, which are required for one of the methods of calculating the angle bending parameters and the periodicity's relating to the torsional barriers.

At this point all the data from files has been loaded and the next step is to enter the run time control parameters.

5.8.2 Run Time Parameters

5.8.2.1 Date

The first of the run time parameters is the date on which the run is initiated. This is entered in the format DDMMYY. The date is used within several of the output file names to keep them distinct and to help identify from which run they were produced. The date is also included in other files for the same reason. This information could be obtained from computer functions but manual input allowed multiple runs to be done in one day, i.e 190490 and 19049B, etc.

5.8.2.2 Previous Run

As stated earlier it is possible to load in the list of which parameters affected which structures from a previous run. If this option is chosen at this point then the user is prompted for the date of the run to be used.

5.8.2.3 Lone Pairs

The next item prompts if lone pair electrons are to be added. Some force fields include lone pairs to increase accuracy by simulating the fact that atoms with lone pairs cannot be considered spherical. It was thought that, if lone pairs were included, that any resultant increase in accuracy might allow the use of a less specific configuration for one or more of the parameter types.

5.8.2.4 Save Parameters

The next item asks if the resultant optimised parameters should be saved to disk. In a normal run of the program the answer should be yes, however when debugging any alteration to the program it is useful to run the program without saving any parameter files. The reason for this is that when debugging it may be necessary

to go through many runs in a short time and if each run saves a new set of parameter files the number of files will soon become unwieldy, especially as in most such cases there will be no point in saving such files.

5.8.2.5 Initial Iterations

As the structures usually start off in the same geometric configuration for each run there will be a difference between these configurations and those which would be expect if the molecule had been minimised for the starting force field. For this reason the program asks for the number of initial iterations. Besides moving the structures towards their minimum energy configurations for the current force field it is also important as the final value of the RMS of the first derivatives is used, as described above, to decide on the criteria at which minimisation's are terminated.

5.8.2.6 Save Files

If initial minimisation's are selected there is then a prompt asking if the files should be saved after the minimisation's. This can be useful if only small changes are being made in the configuration of the force field as it means that next time you will need a smaller number of initial iterations to reach the same point in the minimisation.

This option is rarely used as by keeping the initial structures constant it allows a better comparison of results.

5.8.2.7 Average Non-Bonded

The next question relates specifically to the way that the parameters for non-bonded interactions are implemented and only applies to the A6-B12, A6-B9 and A6-EXP forms of non-bonded interactions (and so in some cases may not appear).

One method of reducing the number of primary parameters in these cases is to specify the values for the interactions between identical atom types and to fill the off-diagonal locations in the arrays with geometric means of the relevant parameters for the two type's involved, i.e.:

$$A_{ij} = \sqrt{A_{ii} A_{jj}} \text{ and } B_{ij} = \sqrt{B_{ii} B_{jj}}$$

If the reply to this question is yes then the program will use this method of reducing the parameter numbers by calculating the off-diagonal parameters. If the answer is no then the program will use the full parameter set.

5.8.2.8 Geometrically Averaged Non-Bonded Parameters

If the reply given in the previous question is no then the user is given the option of starting the non-bonded parameters at the values they would be at if the

averaging of the off-diagonals method was being used.

This option is present to allow easy comparison between the reduced and full situations. If the current force field has been optimised with the reduced force field then by using this option the program can switch to the full representation but start with the same parameter constants produced from the reduced method.

5.8.2.9 Weightings

The next input asks if any weightings are required. As described above in section 5.5.4 there are three contributions towards the measure of closeness. These are the structural properties, the heats of formation and the energy differences between isomers. By entering weightings it is possible to alter what proportion of the final residual factor comes from each of these parts. It is then possible, for example, to highly bias the energy differences between isomers such that it would more rapidly go to zero than if it had a lower weighting.

5.8.2.10 Reminimisation

This option allows the user to select how often reminimisation occurs. Ideally reminimisation should occur at the end of each force field loop. However, reminimisation can take a significant proportion of the time, especially when only one section of the force field is being optimised. In these cases much time can be saved by reminimising less often. Take for example, the situation where a force field is being optimised with 500 iterations for both the initial and the reminimisations and the current situation initially involves 20 force field loops where only the central-two torsion barriers are being optimised. In this case the optimisation of the parameters would take about 5 minutes, whereas the intermediate reminimisations would take 30 minutes. Thus by only reminimising every 5 times round the force field loop the time for the first 20 force field loops will be 220 minutes as opposed to 700 minutes.

5.8.2.11 Force Field Loops

Finally the program prompts for the number of force field loops to be executed. If it were possible to determine the duration for one such loop then it would be feasible to set the program to finish at the required time. This is not generally possible and so the usual practice is to enter a large number of iterations but to break out of the program when it has reached an appropriate position. This is usually after a reminimisation when the most up to date comparison file will have been saved to disk. Because of this approach to terminating the program, all data files that the program saves are updated at the end of each force field loop around the primary parameters.

5.8.3 Program Initialisation

This section relates to the remaining operations that must be executed before the preminimisation can take place.

5.8.3.1 Data From Previous Run

If the user has selected to use the data from a previous run relating to which parameters affect which structures then this file will be loaded at this point.

5.8.3.2 Lone Pairs

If the lone pair option has been selected then the molecular data must be updated accordingly. Each molecule in turn is scanned for the atom types that have lone pair electrons. When such atoms are found then the lone pairs are added automatically in the correct positions with the relevant changes to connectivities and the number of atoms in the molecule. This is done using the procedure called 'add.lone.pairs' (A.1.11.5), this is a variation of the procedure used within the modelling program COMMET to automatically add hydrogen's to a molecule.

5.8.3.3 Bond Multiplicity

Bond multiplicity is a measure of the nature of a bond, i.e. a single bond has a bond multiplicity of 1, a double bond has a bond multiplicity of 2 and a conjugated single bond has a value of 1.1.

Most bonds within the molecule will usually have a standard bond multiplicity. However there are some cases involving possible conjugation, such as a C_{sp2}-C_{sp2} bond, where knowing the atoms involved in a bond is not enough to predict the bond multiplicity. The procedure 'asboml' (A.1.11.1) inspects the atoms, their connectivities and their initial bond lengths and from these decides the multiplicity of each bond in the molecule. This is obviously important as conjugated bonds will have very different parameters to normal bonds.

5.8.3.4 Fixed Parameters

This section looks to see if any of the parameters were fixed during the force field initialisation section. The program then initialises an array for each parameter with a value relating to how long the parameter is fixed, if at all. If fixed permanently, it will have a value of -1, if fixed for n force field loops it will have the value n and if not fixed it will have the value zero. Every time that a force field loop finishes all positive values will be reduced by one until they reach zero, at which point they will cease to be fixed.

5.8.3.5 Distribution of Initial Data

With the necessary arrays initialised the program next sends out a large data block that is passed along the pipeline with each processor in turn taking a copy. The contents of this data block are given in the previous section on MINIMSRS (section 5.6.4.1).

5.8.4 Initial Minimisation

The initial minimisation's are then done in such a manner as to hopefully optimise the load balancing amongst the processors. The method by which this is done is covered in section 5.5.2 on the development of the current load balancing technique.

As each molecule's data packet is returned from the minimisers the program takes all the information and stores it in the relevant arrays. Also at this time a selection of the returned information is displayed. The data shown is the molecule number, the structural residual factor, the final energy, and the final RMS force as well a count of the number of data packets returned. An example of a screen display is as follows:

```
------------------------------------------------------------
INITIAL MINIMISATION
------------------------------------------------------------

STRUC 18 DIF   0.000 ENERGY 17.254 RMS OF 1st DERIVS 9.475E-03  0
STRUC 30 DIF   1.634 ENERGY 11.449 RMS OF 1st DERIVS 1.343E-02  1
STRUC 32 DIF   0.000 ENERGY 12.574 RMS OF 1st DERIVS 8.556E-03  2
STRUC 17 DIF   5.897 ENERGY  9.357 RMS OF 1st DERIVS 7.583E-03  3
STRUC 24 DIF   0.486 ENERGY 26.458 RMS OF 1st DERIVS 8.848E-03  4
STRUC 12 DIF   0.000 ENERGY  9.743 RMS OF 1st DERIVS 9.453E-02  5
STRUC  5 DIF   1.196 ENERGY 10.047 RMS OF 1st DERIVS 8.163E-03  6
STRUC 18 DIF   1.030 ENERGY 22.864 RMS OF 1st DERIVS 1.838E-01  7
STRUC 49 DIF   0.000 ENERGY 11.576 RMS OF 1st DERIVS 9.637E-03  8
```

Once all the data from the initial minimisation's has returned, the program produces its initial COMPARE file that it saves to both the COMPARE directory on the hard disk and to the floppy drive. The copy of the COMPARE file on the floppy is made so that the file can be examined on another machine without the need to terminate the program.

5.8.5 The Force Field Loop

The program will then go round the parameters in the force field optimising each parameter in turn. First it checks to see if the parameter is fixed and if so it proceeds to the next parameter. If a parameter is not fixed then the program must decide on the value of the small shift used in the determination of the derivatives and on the maximum value that the calculated shift can take. The routine that does this is

the only procedure outside of FFCONVRT where the actual parameter type value is used. The reason for this is that certain parameters such as bond angles are proportionally more sensitive to movement than parameters such as non-bonded parameters. The following table contains the multiplication factors used to determine the shift and max shift values.

Table 5.1 Shift factors for parameter types

Parameter Type	Shift factor	Max Shift factor
Bond length	0.000125	0.00025
Bond stretch	0.005	0.02
Non-bonded	0.005	0.02
Bond angle	0.001	0.002
Torsional barrier	0.005	0.02
Out-of-plane bending	0.005	0.02
Calc stretch constant	0.005	0.02
Angle weight factor	0.001	0.002
Angle power	0.005	0.02
Total angle	0.001	0.002

Once these shifts have been decided the molecules affected by that parameter are found by inspecting the appropriate array. How this array is set up is described in section 5.5.3. The load balancing requires that the molecules be sent out to the minimisers in order of decreasing minimisation time. The program contains an array that gives an estimation of this time. A bubble sort routine is invoked to arrange the active molecules in the correct order and then these molecules are sent to the network (section 5.6.4.4). The minimisers then return the information that the control program requires to calculate the various residual factors.

Values are then calculated for all three different types of residual factors for each of the three situations that the minimisers examined (no shift, plus shift and minus shift). For each of these cases the values are combined using the weights to produce a final value for each case. These values are then treated by the Newton-Raphson method to predict the optimum shift.

This value of the optimum shift is returned to the minimisers that originated the required values for this case. Again the total residual factor is calculated and this value is compared with the previous three values used in the calculation of the shift and the one that gives the lowest total residual factor will be selected as the optimum shift. A flag informing the minimisers which is the optimum shift is then sent and the minimisers respond by transmitting the new co-ordinates for this case to all the other worker tasks and the control program.

The reason why the calculated shift is not automatically implemented is that the Newton-Raphson technique assumes the property being optimised is harmonic

about the minimum position. This approximation means that in many cases the optimum shift will be over estimated, so a test is required to check that it does bring about an overall reduction in the combined residual factor.

If the optimum shift is that relating to the calculated shift then a further test is done. In this case the program examines the three individual residual factors in the four cases and if the values for all three of these are at their lowest in the calculated shift case then that parameter is optimised again.

This means that parameters involving reductions in all three measures will be optimised faster than those where the optimisation is a trade off between different residual factors. In order to achieve an even balance between the parameters the program only allows this re-optimisation to occur up to a maximum of 5 times for any given parameter in a single force field loop.

The following is a typical example of the screen display during the course of a parameter optimisation.

```
-----------------------------------------------------------------
ITERATION 44  PARAM 41:  6  0  6  6  2, VALUE : 109.661011
-----------------------------------------------------------------

STRUCTURE: MINUS   0.269550 CENTRE   0.268397 PLUS   0.269568
ENERGY   : MINUS   0.000000 CENTRE   0.000032 PLUS   0.000997
HEAT FORM: MINUS   0.045485 CENTRE   0.044881 PLUS   0.044336
-----------------------------------------------------------
COMBINED : MINUS   0.315036 CENTRE   0.313310 PLUS   0.314901

CALCULATED PARAMETER SHIFT = 4.45162819E-04
SENDING CALCULATED SHIFT
 ALL DATA BACK
ST:  0.268383 EN:  0.000033 HF: 0.044879 COMB: 0.313296

SENDING OPTIMUM SHIFT (NEW)

=================================================================

SENDING NEXT PARAMETER DATA
GETTING DATA BACK
 43/50 [3:3:4:3:3:3:4:1:4:4:3:2:1:1:1:2:1:1:1:1:1:1:1:1] 46:1
```

The following information is displayed on this screen:

The term 'ITERATION' refers to the number of the force field loop. So this example comes from a 44'th force field loop.

PARAM shows the internal parameter number. The internal parameter numbers are assigned in order as the various parameter blocks are loaded in during the initialisation stage. The information following the param is the data used to show what that parameter relates to. In this case the first number is a 6 which relates to

bond angle. The next three numbers are 0,6 and 6 which means that it is the parameter for the bond angle H-C_{sp3}-C_{sp3} and the final value of 2 shows it is to be used only in situations when there are two hydrogen's connected to the central atom.

The VALUE is the current value of the parameter.

The next section relates to the various measures of closeness that have been calculated for the three positions of the parameter used to produce the calculated shift.

Next the shift that is produced by Newton-Raphson optimisation from the combined values is displayed.

The next section will be active when the program is receiving data back from the minimisers. At that time the line under the 'SENDING CALCULATED SHIFT' will show the proportion of data so far returned. For example 6/45 would mean that of the 45 structures affected by a parameter, so far data relating to six of them had been returned. Once all the data is returned the 'ALL DATA BACK' is displayed as in the example.

Once all the data is back it is possible for the control program to calculate the residual factors relating to this case and these are shown on the next line.

With all the data now available the program inspects the four values for the combined residual factor and selects the lowest of these to be the optimum shift. This case is sent to the minimisers and the screen acknowledges what the optimum shift was and the fact that it has been sent. If the situation arises where all the measures are lower for the calculated shift then this line will be followed by one saying 'ALL VALUES LOWER, OPTIMISING AGAIN'.

The section following the double line across the screen relates to the next parameter. The screen is arranged in this manner to allow for the visual inspection of the data from the previous optimisation as well as showing the current state of the optimisation of the next parameter. The first two lines in this section show that the data for the next parameter has been sent and that the returning data is being received.

The final line shows the progress of the minimisation's for the next parameter. The first two figures relate to the number of data blocks received as compared to the number of structures sent. Both figures will go up initially as data blocks are received from the minimisers and new structure blocks are sent out to replace them until all required structures have been sent. This stage will finish when all the associated data packets are returned.

The next part of the line shows the distribution of the structures in the network. Each number within the square brackets shows the number of structures that are being run on each processor in the pipeline. So for the above example we can see that the first processor has 3 structures active, the second also has 3 and the third

has 4. Inspection of the distribution gives some idea of how much of an effect a particular parameter is having. If a parameter is causing some molecules to have to undergo a large number of minimisation's then the distribution will be uneven. If all molecules are minimised sufficiently within a few iterations then we can expect a more even distribution.

The last two figures on the line relate to a molecule and the processor it was running on for the last set of received data.

As the values for the various residual factors are comparable from parameter to parameter it is possible to produce a graphical display showing how the values are changing during the course of the optimisation.

Initially this display is done using a Digisolve VGP64 graphics processor that has been adapted to accept its input through a transputer link from the control program. The form of the graph plotted was that all 4 measures are scaled so as to use the full range of the screen and the scale given related to the total residual factor. A mark is made on the x axis to indicate the end of each force field loop (Figure 5.19).

Towards the end of the project a Datapath card was used. This has a greater resolution and has allowed the addition of new features. The current display shows the scales for all four residual factors (total, structure, isomer energy and heat of formation). it also contains a histogram type display showing the size, relative to the maximum shift, of the last applied shift for each parameter (figure 5.20).

Figure 5.19 Layout of Screen Output Using VGP64

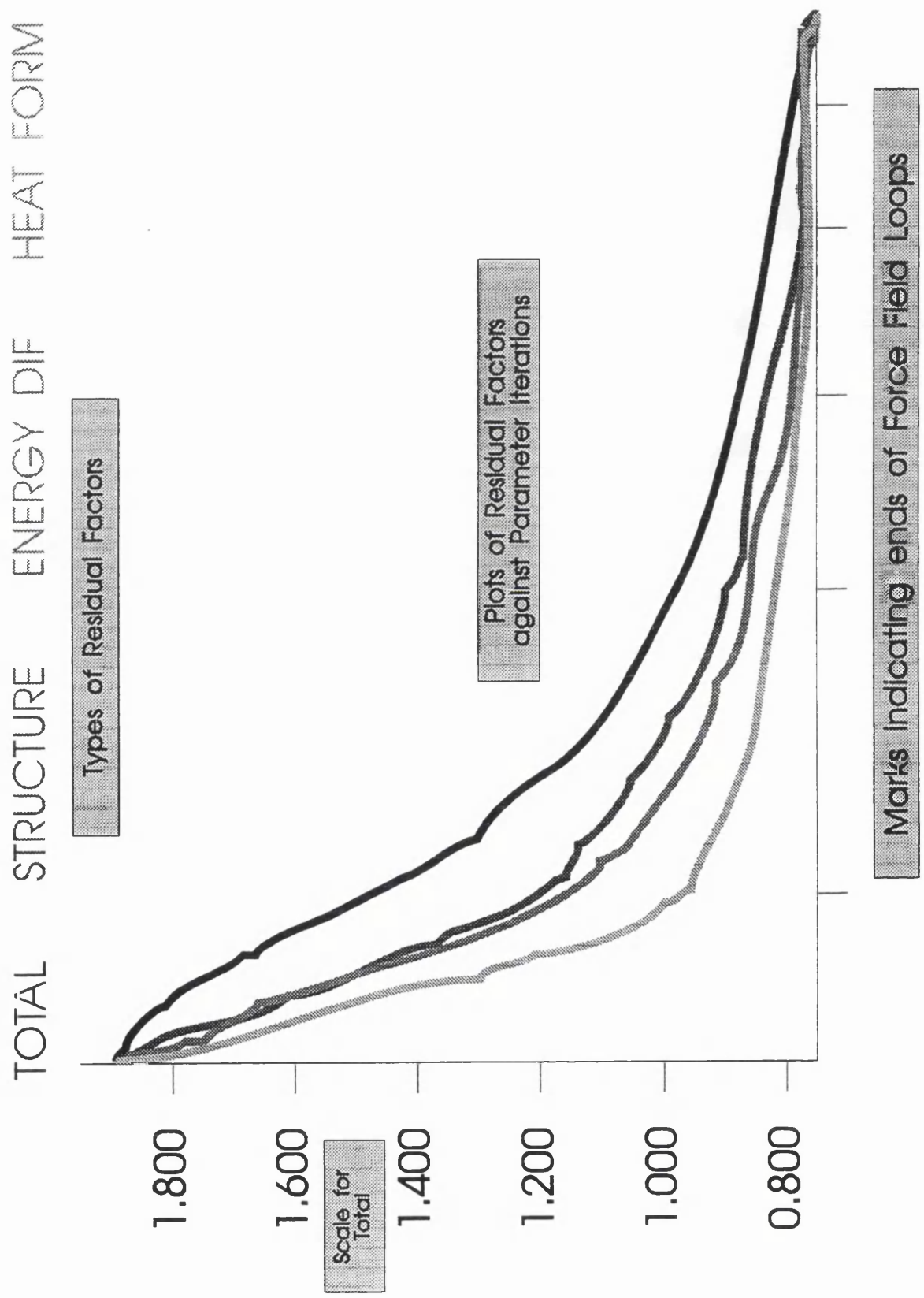
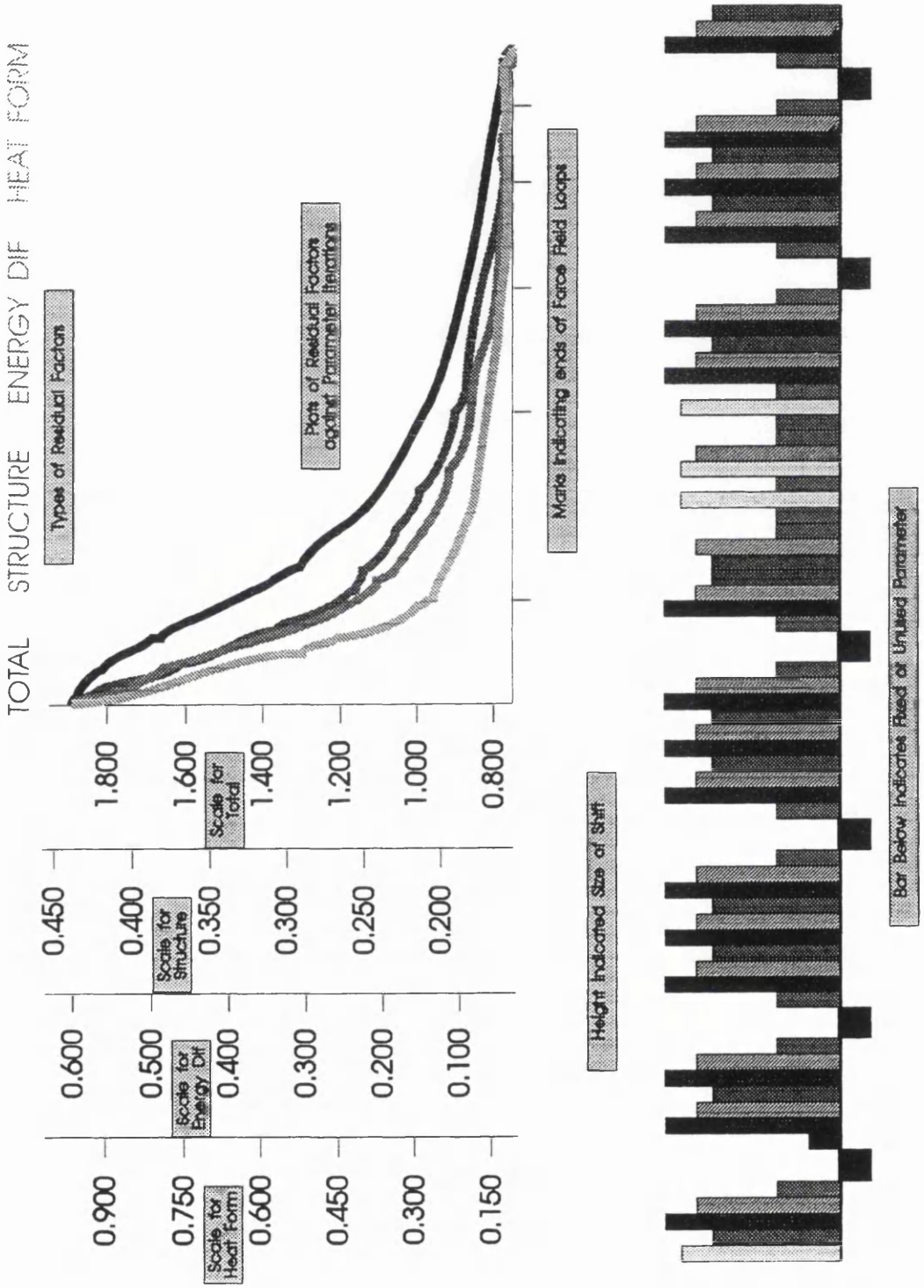


Figure 5.20 Layout of Screen Output Using Datapath Card



At the same time that the graph is plotted the program saves a file containing all the values that are being plotted so that the graphs can be reproduced using other systems if required.

At the end of a force field loop the program will, if required, save files to the current data set containing the current parameter values. At the same time an associated information file is saved for each parameter block. An example of such an information file can be found in appendix C.

5.8.6 Reminimisation

After the end of each force field loop all structures can be reminimised for the current force field. The number of iterations is the same as that used in the initial preminimisation. This is done to make sure that at the beginning of the next force field loop all the structures will be in their low energy positions and to produce a new value for the RMS of the first derivatives for use in the minimisation criteria. The approach used for reminimisation is identical to that used in the initial minimisation. Similarly after minimisation and update of data (including the molecular co-ordinates) a COMPARE file is produced which shows the closeness of the structural properties and the various energy properties as well as including other data relating to the associated force field loop. A typical COMPARE file can be seen in appendix B.

5.8.7 Termination

The above processes will be run until the selected number of force field loops have been completed. Once this has happened the program will terminate after the final COMPARE file has been written and the TDS system will be restored to the root transputer.

5.9 OPTPARAM Implementation Problems

This section discusses some of the problems involved in creating a large occam program running on a network and some of the techniques used during the debugging.

A significant problem was that because the vast majority of the program was written from scratch the code contained several coding errors. Some of these bugs were trivial, but others had a large effect on the results. This meant that comparisons could then only be made between runs using the same data and the same, hopefully debugged, version of the program.

The difficulties involved in debugging the program fall into two categories. The first is that the program is trying to produce a new force field for each

configuration so there are no standard values that the final force field data can be compared to. The best that can be done is to examine the parameters and compare them with values from other force fields. If the values are very different there are two possible reasons, a bug in the program or an insufficient data set.

The second problem relates to the difficulty of debugging a parallel processing network, especially in a MIMD machine such as a transputer network. This is added to by the fact that some of the bugs in OPTPARAM tended to show up only after many hours of processing and when the initial load balancing was used any problem was likely to occur on a random processor.

Until the occam debugger became available in the most recent release of the TDS it was almost impossible to debug these situations. At most, attempts could be made by sending debugging information down links that were connected to other PC's, so that the debugging data appeared on the screen.

Even with the debugger, the situation is not perfect, as when the debugger traces an error, be it a coding error or a deliberate debugging STOP, it finds the location in the code but not the associated work space. This means that the data being acted upon at the time could not be examined, thus causing a severe drawback to the debugging.

There are several ways of getting round this problem. One is to replace a debugging STOP by a very large time delay using a TIMER. In this case the program is terminated manually and the debugger shows the code pointer and workspace for the process waiting on the timer, so allowing the data at that point to be investigated.

The other option does not require the addition of this sort of loop but requires a certain degree of exploration by the programmer to find the correct workspace. Any process is likely to be linked to another process by channels and these channels can be inspected to find the code pointer and workspace associated with them. Thus by inspecting these values for the channel connected to the process on which the problem occurred it is possible to locate the workspace pointer for this section. To get into this network of channels the code pointer and the workspace pointer for any channel must be found. When a processor is waiting on a link it will be detected by the debugger and the associated code pointer and workspace pointer can be obtained. This allows access to the data of the control process for each processor. By inspection of the various internal channels to and from the worker tasks it is then possible to access the data on each task.

The only remaining problem is to know on which task the problem occurred. This is not helped by the fact that identical processes have the same code pointer but different workspace pointers. So even though the initial debugging trace gives us the

code pointer this is of no use in determining which minimising task the problem occurred on. Unless the identification number of the affected structure is known the only method is to try and find out which minimising process failed by examining the control process data. This is simple when the minimisers are working one after the other, as in the small shifts for the optimising situation, because the problem is located on the last minimising task allocated which can simply be found by examining the correct variable on the control task.

The final method for overcoming the short comings of the debugger takes advantage of the fact that any channel expecting data can be identified. This identification includes both the code and the workspace pointer, thus allowing the code in the vicinity of the problem to be inspected properly.

If the program crashes at a STOP then the STOP can be replaced by:

```
CHAN OF ANY debug :
INT tmp :
PAR
  SEQ
    STOP
    debug ! tmp
  debug ? tmp
```

Thus the STOP occurs before the value of tmp can be sent down the debug channel and thus the statement waiting for tmp can be identified using the debugger.

A similar approach can be used for errors that occur at some other statement, for example, a divide by zero arithmetic error. In this case it is possible to place the whole process under consideration in parallel with the debugging channel command, i.e. going form code of the form:

```
... section without error
... section with error
... section without error
```

to:

```
... section without error
CHAN OF ANY debug :
INT tmp :
PAR
  SEQ
    ... section with error
    debug ! tmp
  debug ? tmp
... section without error
```

This has the advantage that the code will act exactly as before until the time the code is executed when the error occurs.

The only draw back to this approach can occur when debugging code in libraries. This is because when errors occur in libraries the relevant folds are opened

to the point at which the error occurred but it is not possible to inspect other folds. Thus in the above code we could locate the waiting debug channel statement but would then be unable to inspect the code fold above it. This can of course be alleviated by removing most of the folds from the libraries to leave a single fold.

Thus we can see that the development and debugging of a large occam program can take, and in fact did take, a substantial time.

5.10 References

1 A.L.Verma, U.F.Murphy and H.J.J.Bernstein, *J. Chem. Phys.*, 1974, **60**, 1540.

2 Thermochemical Data of Organic Compounds, J.B.Pedley, R.D.Naylor and S.P.Kirby, Chapman and Hall Ltd., 1986, ISBN 0-412-27100-1.

3 D.N.J.White, J.N.Ruddock and P.R.Edgington, *Molecular Simulation*, 1989, **3**, 71.

CHAPTER 6

Data Selection and Preparation

6.1 Introduction

Before OPTPARAM could be run, it was necessary to produce all the associated data files. These data files are :

Structure files
An isomeric energy difference and structure list file
A constants file
Initial parameter files

Of these files, most work is involved in the construction of the data files relating to the structure. The reason for this is that these files contain:

i) The initial molecular structure, including the charges on the atoms if the monopole charge interaction is used.
ii) Any available structural experimental information (e.g. bond lengths) including experimental errors.
iii) The experimentally determined heat of formation data with its related error, if both are known.
iv) The number of each enthalpic increment group required to calculate an estimate of the heat of formation from the steric energy. These enthalpic increments are required only if the experimental heat of formation is given.

6.2 Structure data files

Before any structures could be selected and processed to produce the required input files, a choice had to be made as to the source of structural data to be used. The choice is between structures that have been determined in the crystalline phase, mainly by x-ray diffraction and the vapour phase, mainly by electron diffraction.

The main advantages of x-ray diffraction are that there exists a large number of experimentally determined structures[1] and that the position of non-hydrogen atoms in such structures are relatively well defined.

A drawback to using x-ray diffraction data is that the positions of the hydrogens are not so well defined. This means that the effect of any change to a parameter used in an interaction involving one or more hydrogens can only be judged by its influence on the positions of the non-hydrogen atoms. Also to reduce the computation time the optimisation is carried out on a single molecule, a situation which is obviously inconsistent with the x-ray case where crystal packing effects can be significant, especially on the more easily distorted variables such as torsion angle

which can be the hardest to reproduce using molecular mechanics.

In electron diffraction the internal co-ordinates involving hydrogens can be determined, although with a greater degree of uncertainty compared to non-hydrogren cases. Electron diffraction is also performed on gaseous samples, thus removing the problems concerning crystal packing.

The disadvantages of using electron diffraction are that the number of molecules whose structures has been determined is much lower and that the uncertainties in the internal co-ordinates excluding hydrogens are greater. Also the papers on structures determined by electron diffraction tended to only contain those internal co-ordinates which the author considered important. In X-ray studies there is usually a similar list given but the cartesian co-ordinates are also given, allowing the full list of lengths, angles and torsion angles to be reconstructed.

After considering these points it was decided at least initially to use structures determined by electron diffraction as the gaseous conditions under which structures are determined by this method most closely match the single molecules used in the optimisation. When further force force field development requires optimising parameters for atom types, such as metal ions, not adequately covered by electron diffraction then the suitability of other methods, such as X-ray difraction will have to be reconsidered.

After deciding to use electron diffraction data, the next step in the construction of a suitable database of structures was to select the molecules that would be used in the optimisation. To do this, an initial survey was undertaken of published structures determined by electron diffraction. Because the force field is to be expanded later to cover other atom types, this survey included molecules containing more atom types than those required initially. At this stage just the structures were noted, along with information relating to the existence of isomers and, if present, whether the energy differences were given. An example of this initial data sheet is give in appendix G.

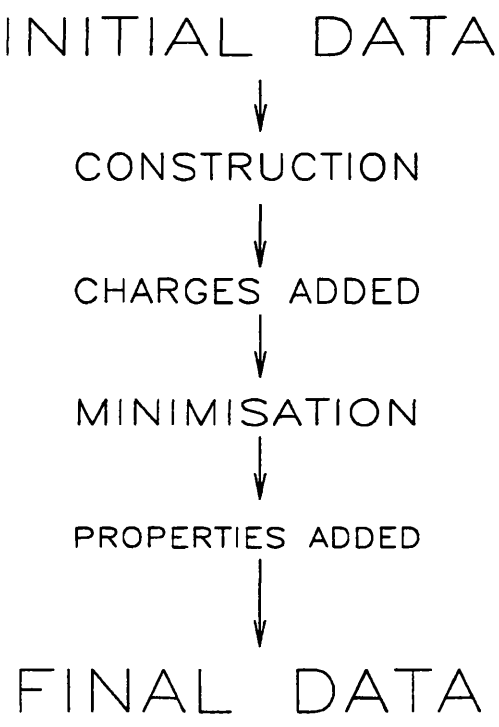
Once the list of structures was completed it was then possible to select which molecules should be used in the optimisation. The aim at this point is to select enough structures to cover all possiblities, while trying to keep the total number of structures as small as possible.

This is a relatively simple task for the alkene force field, but for other atom types there are problems in either choosing or obtaining enough data. The task for the alkenes case was simplified by having access to the structures used in the production of the White Bovill Alkene Force Field (WBFF)[2]. These formed the core of the structure set for the investigation to which were added selected structures that had been determined since the the WBFF was produced.

Once a list of the structures to be used had been constructed, the next step was to return to the literature and make up a more detailed data sheet (see appendix F) for each case. This sheet contained the full name of the molecule, its structure, the geometric properties from the experiment data and the energy difference between isomers, if any.

Each structure was then processed to produce OPTPARAM usable files. (figure 6.1).

Figure 6.1 Order of Data Processing



More specifically these steps are:

Construction

As no Cartesian co-ordinates were available for the structures that had been determined by electron diffraction the molecules were constructed using the PC-CHEMMOD molecular modelling system. When a molecule was constructed its torsion angles were set to those taken for the experimental paper. The other structural properties were not set to the experimental values as the structure is minimised to remove close non-bonded contacts. A molecule can either be constructed from scratch using the programs quick build option or produced by loading an already constructed molecule which could then be simply altered to produce the new structure. For example, the cis but-2-ene molecule could be made from scratch but trans but-2-ene could then simply be made by altering the central torsion angle.

Charge

As discussed above in Chapter 3, there is a choice in the way that the charges can be implemented in molecular mechanics.

This choice is between point charges and dipoles and it is at this stage that a decision must be made as to which system will be used. For the current force field under development it was decided to use the point charge approach. There are several reasons why this approach was chosen :

1) The final force field is eventually to be used in the study of molecular interactions with ionic crystalline lattices and it is far easier to represent their charge distributions using point charges. Using dipoles for the non-crystalline, organic part of such a system would mean having to calculate both dipole-dipole and dipole-monopole interactions.
2) By using an available partial charge calculation program[3], it is possible to reduce the number of parameters that need to be optimised. This is because, in the dipole case, the various dipole moment values relating to bonds between different atom types would have to be treated as optimisable parameters.
3) In the minimising program itself, the calculation of the energy using point charges takes less time, as only the squares of the distances need to be found from the molecules, whereas, in the dipole case various angles and their related trigonometric function values need to be calculated.

To calculate the point charges, the CHARGE2[3] program was used which uses a parameterised σ Scheme approach. The program was adapted from the original so that it could both read and write PC-CHEMMOD style data files.

Other force fields have been optimised using more standard methods for the calculation of atom charges, but a comparison of these methods to CHARGE2 has shown that CHARGE2 is consistently more accurate[3]. As the aim of the optimisation is to produce as accurate a force field as possible it was decided to use the CHARGE2 rather than a more standard but less accurate method such as that suggested by Gasteiger[4].

Minimisation

As described in the section 2.4.5 a Newton-Raphson minimisation can become unstable if it starts in a situation where two or more atoms have van der

Waals surfaces which overlap. The technique developed to overcome this problem results in an increase in the time per iteration and is not applied within OPTPARAM, where speed is essential. Instead the molecule is preminimised to remove any such initial overlap. This was done using a version of the CHEMMIN[5] minimiser which does apply the aforementioned technique.

Property Setting

The structural properties (e.g. Bond length) were entered using a program called PROSET, which is based on the PC-CHEMMOD system.

When run, the PROSET program initially prompts for the name of the file to load. This will be the name of the file that was produced by the minimisation of the molecule in the previous stage. The selected file is then loaded and the molecule displayed. The molecule is transformed before being displayed so that the average deviation from the XY plane is at a minimum. This is done as it usually gives a 'best view' of the molecule and so allows for easier selection of atoms.

Once the molecule is displayed it is then possible to enter the experimentally determined structural properties, if any, for that molecule. This is done as follows :

1) The type of property is selected from a menu.
2) The atoms relating to the selected property are selected either by clicking the mouse on the required atom or by pulling down a 'keypad' and then entering the atom number numerically.
3) The value and error for the property is entered using the 'keypad' on the screen.

Once all the structural properties have been entered 'finish' is selected from the menu and the program goes on to the input of the heat of formation data.

The program prompts first for the heat of formation and then for the experimental error on this figure. If no heat of formation data is available then both these values would be entered as zero.

When the heat of formation is being used the program procedes to analyse the molecule to determine how many of each sort of geometric enthalpy increment will be needed to calculate the correct heat of formation. Once the heat of formation data has been entered, the structure and the entered data is saved in a form that can be read by the OPTPARAM program (see appendix E for an example of this type of file).

6.3 Structure List File

Once the individual structure files have been constructed the next step is to produce a file linking these structures together to produce a 'structure set'. Because the data relating to the energy difference between isomers relates to more than one

structure it must also be declared at this level. The Structure List file then consists of two connected sections. The first of these is a list of the structure files for the current structure set. The second uses the positions of these structures in the list to specify the isomers to be used in the isomeric energy difference comparison.

The first section then, is simply a list of the full file names for the structures used. To inform the OPTPARAM program that this section has finished an 'END' is placed at the end of this list. For example :

```
C:\OPTPARAM\MOLDATA\ISBUTENE.DAT
C:\OPTPARAM\MOLDATA\CYHPTONE.DAT
END
```

The second section starts with the number of energy differences. A list of data is then given, with each line containing the numbers of the structures involved followed by the experimental energy difference and experimental error. For example, the file could end as follows:

```
3
   1   2    0.834 0.143
   3   4    1.130 0.380
   5   6    3.050 0.150
```

An example of a small STRUCTUR.LST file is given in appendix D.

6.4 Constants

The constants relating to electronegativity and the periodicity of bonds were extracted from the data contained within the CHEMMIN[5] minimiser. This data was then saved in a file which could then be read in by the OPTPARAM program.

6.5 Initial Parameter Files

The initial parameter values at the start of a run will have an effect both on the final optimised results and the rate at which those results are obtained. With some parameters such as the bond length the only effect of having an initial value far from the optimum will be that its optimisation will take longer because more steps will be required before the parameter can come close to its optimum value. At the other extreme, if the relative values of some of the torsional barriers are initialised incorrectly, then it is possible that this will lead to an initial deformation of the molecule well away from its experimental structure. In cases like rings this could mean that, even if that parameter were to be optimised towards an optimum value, the ring may stay in the twisted position.

The initial data was obtained in several ways.

Other Force Fields

When available, the initial values were taken from other currently available force fields. For example, the initial parameters relating to the atom types in the COSMIC non-bonded option were taken from the original COSMIC[6] paper.

In cases where the degree of complexity is increased an estimation for the starting values can be obtained from a less complex force field and would result in several parameters starting with the same value. Take, for example, the most complex situation where a reference bond angle parameter is defined by both the atoms in the angle and the degree of hydrogenation. If no parameter values are available "off the shelf" for this situation, then all the parameters relating to a given set of atoms regardless of the degree of hydrogenation can be given the value for that angle from a force field where the parameter is defined by the atoms in the angle only. This then gives an initial situation where the force field effectively mimics the atoms in the angle only case. The difference is that optimising from this point allows these various values to diverge.

From Other Equations for the Same Interaction

This method is mainly used in the calculation of initial parameters for the non-bonded interactions. In this method the values for the non-bonded parameters are taken from a force field with one equation and used to calculate the values of r_0 and e for each required combination of atom types. r_0 is the distance where the interaction energy is e, its greatest value. These values can then be used to calculate the required parameters for a different form of the interaction equation.

For example, the interaction used in the CHEMMIN minimiser is an A6-B12 interaction where the values A6 and B12 for the H-H interaction are:

$A6 = 72.9$ Kcal mol^{-1} $Å^6$ $B12 = 26572.0$ Kcal mol^{-1} $Å^{12}$

Now these values can be related to r_0 and ε by the following equations:

$A6 = -2\varepsilon r_0^6$ (1a) and $B12 = -\varepsilon r_0^{12}$ (1b)

From these we get that:

$\varepsilon = -A6 / 4 \times B12$ and $r_0 = (2 \times B12/A6)^{-6}$

This gives the values:

$\varepsilon = -0.05$ Kcal mol^{-1} and $r_0 = 3.0Å$

Just as equations 1a and 1b relate these values to the parameters for a 6-12 interaction, so similar equations can be produced for a 6-9 interaction:

$A6 = -2\varepsilon r_0^6$ (2a) and $B9 = -\varepsilon r_0^9$ (2b)

Putting in the previously calculated values of e and r_0 into the equations 2a and 2b gives an identical value for the A6 parameter and a value for the B9 parameter of 984.15 Kcal mol^{-1} $Å^9$.

Other Methods

If the required parameters cannot be obtained from either of the above approaches then some other approach must be used. For bond length, averaging of the values of the length in the current data set will usually give a good starting value. If an explicit value for the bond strength constant is required then an initial value can be calculated by employing the formula used in the calculated bond strength option and then optimised from there. For other parameter types the main approach left is that of educated guess work and comparison with similar systems.

6.6 Debugging Data files

Once all the data files have been prepared it is possible to start a run of the program. Initially, when extra data has been added it is necessary to check that this new data had been entered and processed properly. This is most important for torsion angles in molecules with symmetry as it is possible to have the correct configuration but have the sign of a torsion angle inverted. At this stage it is also usually possible to find any typographical mistakes made in entering the data and to check that the part of the PROSET program for calculating the numbers of each type of geometric enthalpy increments is working properly.

Most of these mistakes can be found by a trial run of the program as the properties with the greatest differences from the experimental values will usually be found to be those in error.

After all the input data has been thoroughly debugged it is then possible to start valid program runs.

6.7 Group Enthapy Increments

Initially the group enthalpy increments related to specific groups, for example, a C_{sp3} with 2 hydrogens attached to other carbons. This system worked well in the case of non-conjugated alkenes as only 7 groups are required. When the number of atom types used is increased it results in a large increase in the number of

groups required, each of which must be defined so that it can be identified when the molecule is scanned. This involves explicitly defining a unique number to each option.

To overcome this problem a system of group enthalpy increments was implemented relating only to the bonds in the molecule. Thus each possible bond has a number relating to it which can be simply calculated from the atom types involved in the bond and the degree on conjugation. The equations used to calculate these numbers are as follows:

Non Conjugated

Lowest atom type * numtyp + higher atom type

Conjugated

Lowest conjugated type x numcnj
+ higher conjugated type
+ (numtyp + 1) x numtyp

Where numtyp is the maximum number of atom types likely to be used, in the current situation this is set at 40 (appendix H). numncj is the number of these structures which can take part in a conjugate bond.

At the time that the changeover was made from the initial group enthalpy method to the bond related enthalpy increments a note was made of the effect on the heat of formation residual factor produced in OPTPARAM. It was found that the change had minimal effect on this measure. This implies that the bond related method is just as valid a method of calculating heats of formation as the initial group related method.

6.8 References

1 F.H.Allen, O.Kennard and R.Taylor , *Acc. Chem. Res.*, 1983, **16**, 146.

2 D.N.J.White and M.J.Bovill, *J. Chem. Soc. Perkin* **II**, 1977, 1610.

3 CHARGE2, R.J.Abraham, G.H.Grant, B Hudson and P.E.Smith, University of Liverpool, Liverpool, UK.

4 J.Gasteiger and M.Marsili, *Tetrahedron*, 1980, **36**, 3219.

5 D.N.J.White, J.N.Ruddock and P.R.Edgington, *Molecular Simulation*, 1989, **3**, 71.

6 J.G.Vinter, A.Davis and M.R.Saunders, *J.Comput.-Aided Mol. Design*, 1987, **1**, 31.

CHAPTER 7

Application of OPTPARAM to Evaluate Methods of Parameter Reduction and the Production of an Alkene and an Alkene, Halogen and Oxygen Force Field

7.1 Introduction

The program OPTPARAM was used to evaluate various combinations of parameter reduction techniques to find which combination would give the best balance between the total number of parameters and the correlation to experimental data.

The program was used on two types of structure sets. The first of these was an alkene structure set, as any general molecular mechanics force field will require a good alkene force field as its base. As well as this, to give a broader test of the different options for the interaction types, the structure set was expanded to include compounds that, in addition, contained Fluorine, Chlorine, Bromine and double bonded Oxygen to produce a new structure set. This second structure set will be referred to as the expanded structure set. This structure set was itself extended by the addition of extra data, so any reference to this structure set will be specified as being either the first or second expanded structure set. The structures used in all these structure sets are given later in this chapter.

Chronologically the first structure set to be investigated was the alkene structure set. Next the first expanded structure set was investigated, followed by the second expanded structure set. During the course of these optimisations several errors in the program became known and so finally several of the alkene structure set optimisations were re-investigated.

The approach used was to keep fixed the other parameter groups (see section 5.8.1 for an explanation of the different groups) and vary only the form of the parameter group under investigation at the time. This allows for the investigation of the effectiveness of each method in turn.

With the benefit of this information it was possible to select, for each case, the optimum method by balancing the results obtained against the number of parameters, taking into account not only the number of parameters that are required for the current structure set, but also the number that would be required if a large number of atom types were to be used. The reasoning behind this is that the final aim must be to produce a complete force field that is usable for many atom types and that some options that lead to a manageable number of parameters for the restricted case under current investigation may be seen to be unwieldy when applied to a structure set with many more atom types.

Using the selected 'optimum' methods the force field was then re-optimised and the results from this final force field are compared with the previous optimisations to see what effect the combination of the chosen methods has had.

Finally the parameters for this final optimisation are given and their generality

is discussed.

7.2 Preliminary Considerations

Before examining the results and what can be deduced from them it is necessary to investigate several points that were not covered in the discussion of the program implementation in chapter 5. These points relate not to the actual implementation of the OPTPARAM program but to such things as the choice of weightings and situations that occur during actual force field parameter optimisations.

7.2.1 Incomplete Optimisation and Optimisation Progression

It is one of the major problems of empirical optimisation techniques, such as that used in these experiments, that though there is a progression towards an optimum position it is seldom possible for complete optimisation to occur. In the current case complete optimisation would occur only if no parameter was altered during the course of a complete force field loop. In all the optimisations that took place, including the final optimisation involving 100 iterations using the second expanded structure set, this situation never occurred.

A further problem is raised by the fact that the number of iterations applied during each evaluation of the different methods under comparison were different, and even had they been the same, there is no reason why they should be comparable. For example, if two methods of determining the non-bonded interactions were to be compared and the initial parameters where far closer to the 'optimum' values for the first case, then it is likely that these parameters are going to tend to stay closer to their 'optimum' values ahead of the second case for an equal number of iterations.

The information available at the termination of an optimisation run was the graph and other data displayed showing the progress of the optimisation. As far a possible each run was continued until the graph showing the movement of the residual factors was relatively flat and in the later cases, where the relative size of shifts was shown, that these shifts should be small. The problem, as with most forms of such optimisations, is always one of diminishing returns.

For example, Figures 7.1 and 7.2 show the change in all the residual factors for two typical program runs. The optimisations took place on the alkene structure set for figure 7.1 and on the second expanded structure set for figure 7.2, both runs had no fixed parameters and the weightings used in both cases were as follows:

Structure	: 5
Energy Difference	: 20
Heat of Formation	: 1

Figure 7.1 Typical Residual Factor Plots for the Alkene Set

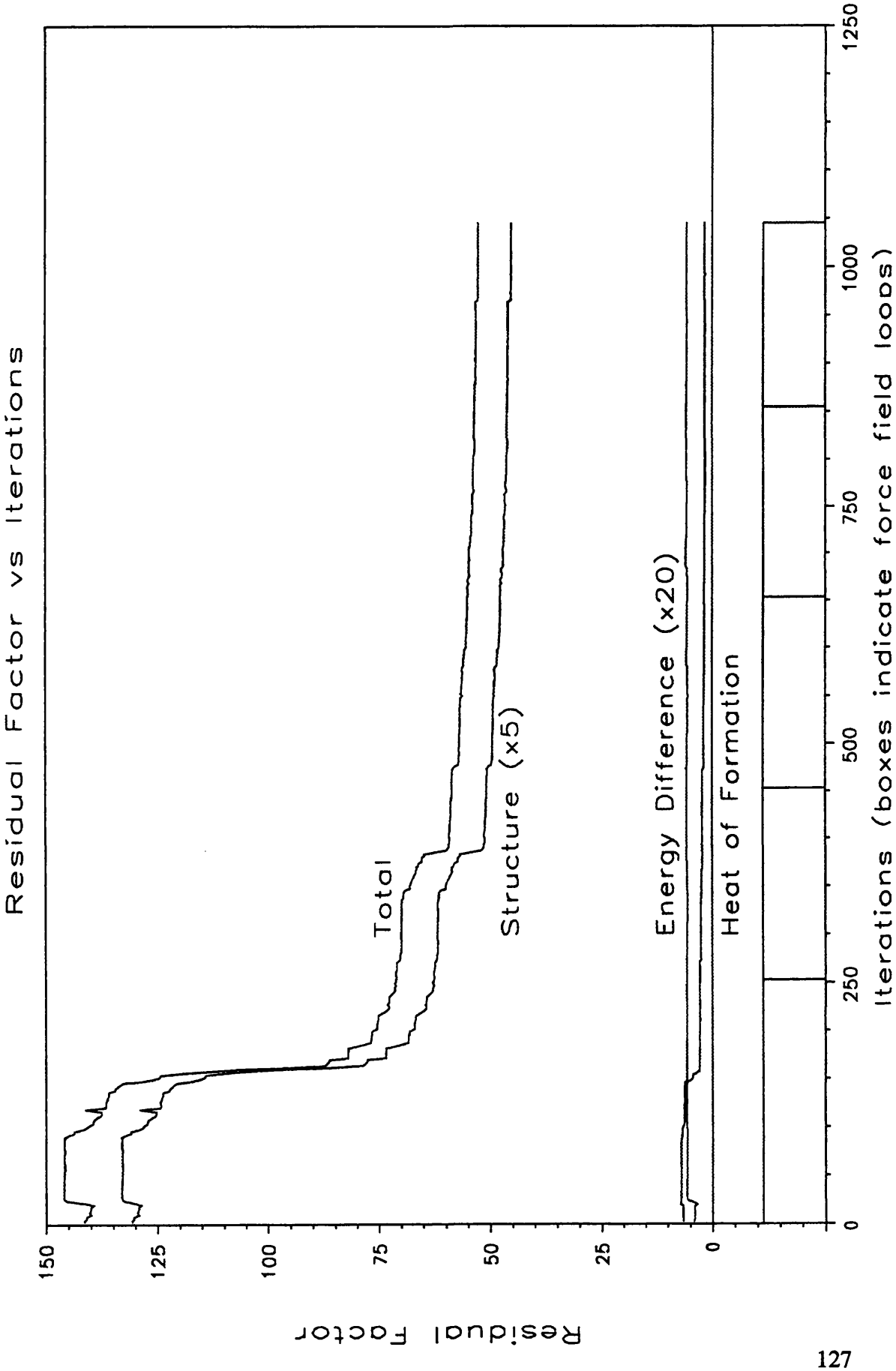
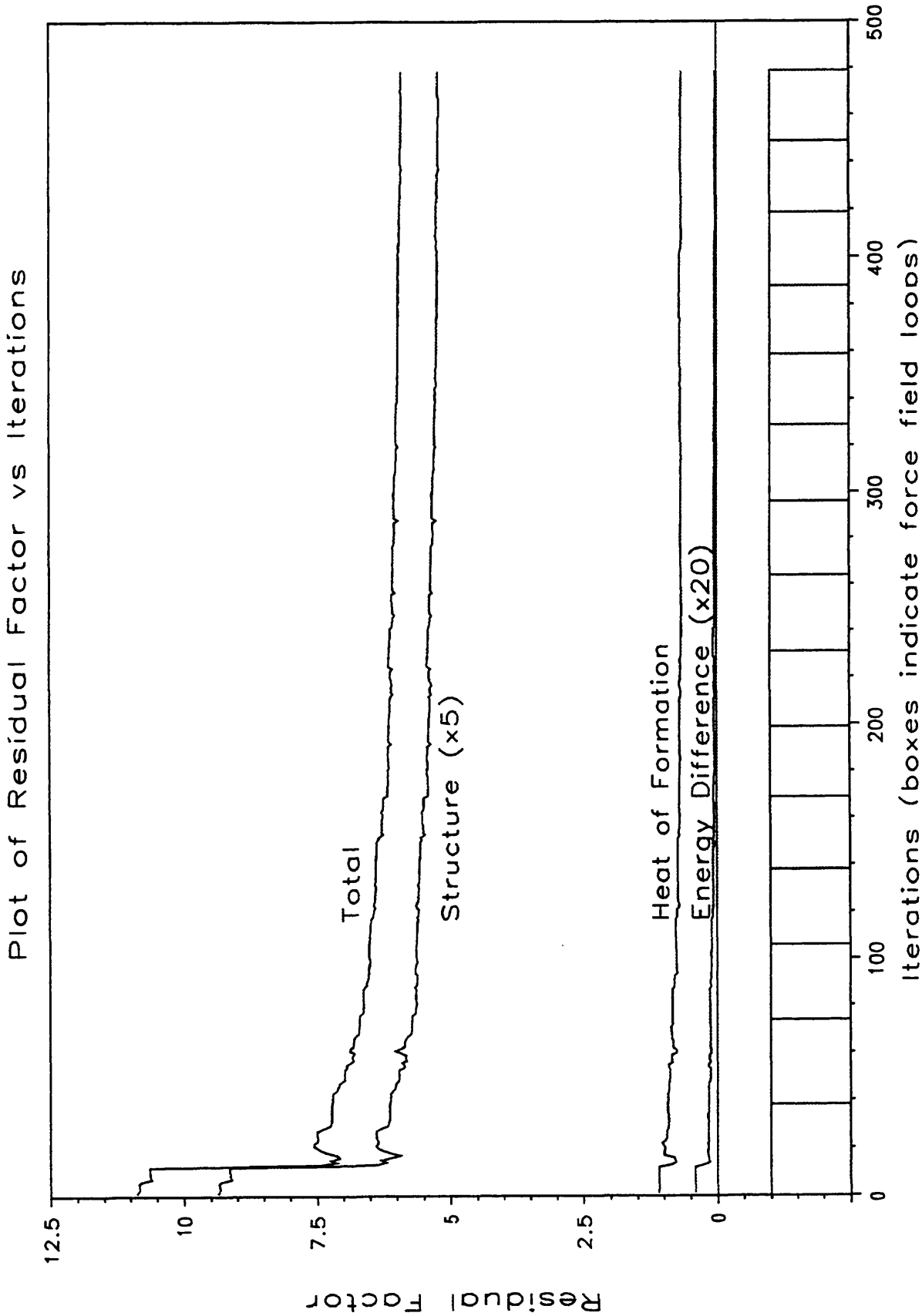


Figure 7.2 Typical Residual Factor Plots for the 2nd Expanded Set



The force fields being optimised in these cases had the following configurations:

	Figure 7.1	Figure 7.2
Bond Length:	Full	Full
Bond Stretch:	Calculated	Calculated
Non-Bonded:	Cosmic[1]	A6 - EXP
Angle Values:	Calculated	Full
Angle Bend:	Central Only	Full
Torsion:	Central Two	Central Two
Out of Plane:	Implemented	Implemented

The values plotted in the graphs include the weighting factors, so it can be seen that the curve for the total residual factor comes from the combination of the other three curves. The boxes at the bottom represent the force field loops, Thus for the cases shown in figures 7.1 and 7.2 a total of 15 and 5 force field loops were completed respectively.

These graphs then, show the movement of the residual factors with iterations. In addition, during the course of some of the optimisations the current values of all the parameters were saved at the end of each iteration. This data can be manipulated to produce graphs that show how the parameters have moved during the optimisation. Figures 7.3 and 7.4 show the movement of the parameters during the course of the optimisations shown in figures 7.1 and 7.2. Rather than display the movement of all the parameters individually, the parameters have been grouped into their section types. These section types are :

Bond length	Bond stretch	Non-Bonded
Valance angle	Angle bend	Torsion barrier
Out of plane bending		

For each force field loop the absolute percentage difference between the final and current value of each parameter was calculated. An average was then calculated for each parameter section, as shown above, and these values were then plotted.

Comparison of these curves against the residual factor movements shows that even when there is only a small change in the residual factor compared to previous force field loops there can still be significant shifts in parameter values. This can be explained by considering a residual factor' surface similar to a potential energy surface. As the position on this surface approaches a valley associated with the 'optimum' force field, then similar sized parameter displacements are likely to result in smaller decreases in the residual factor.

Figure 7.3 Typical Parameter Movements For Alkene Set

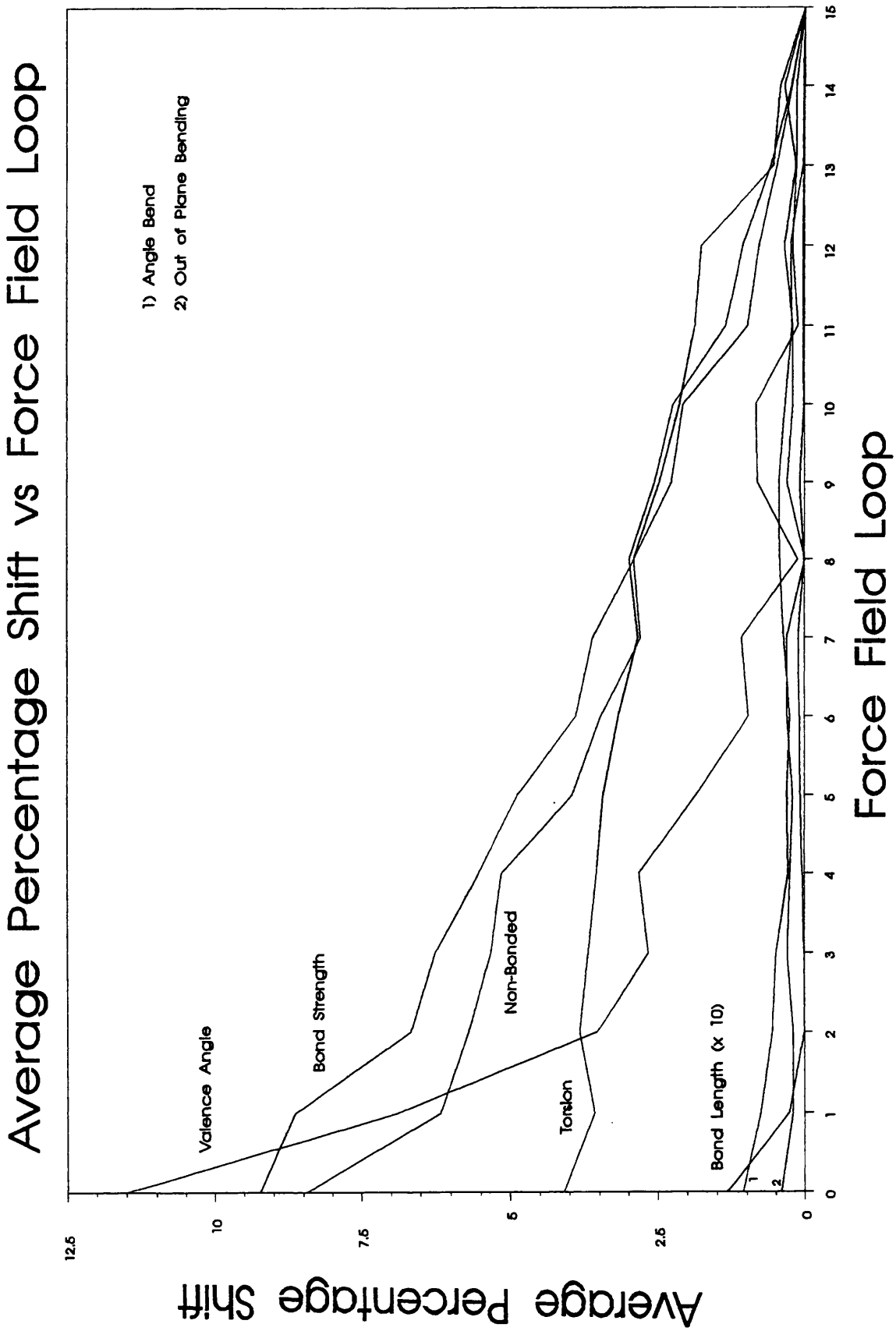
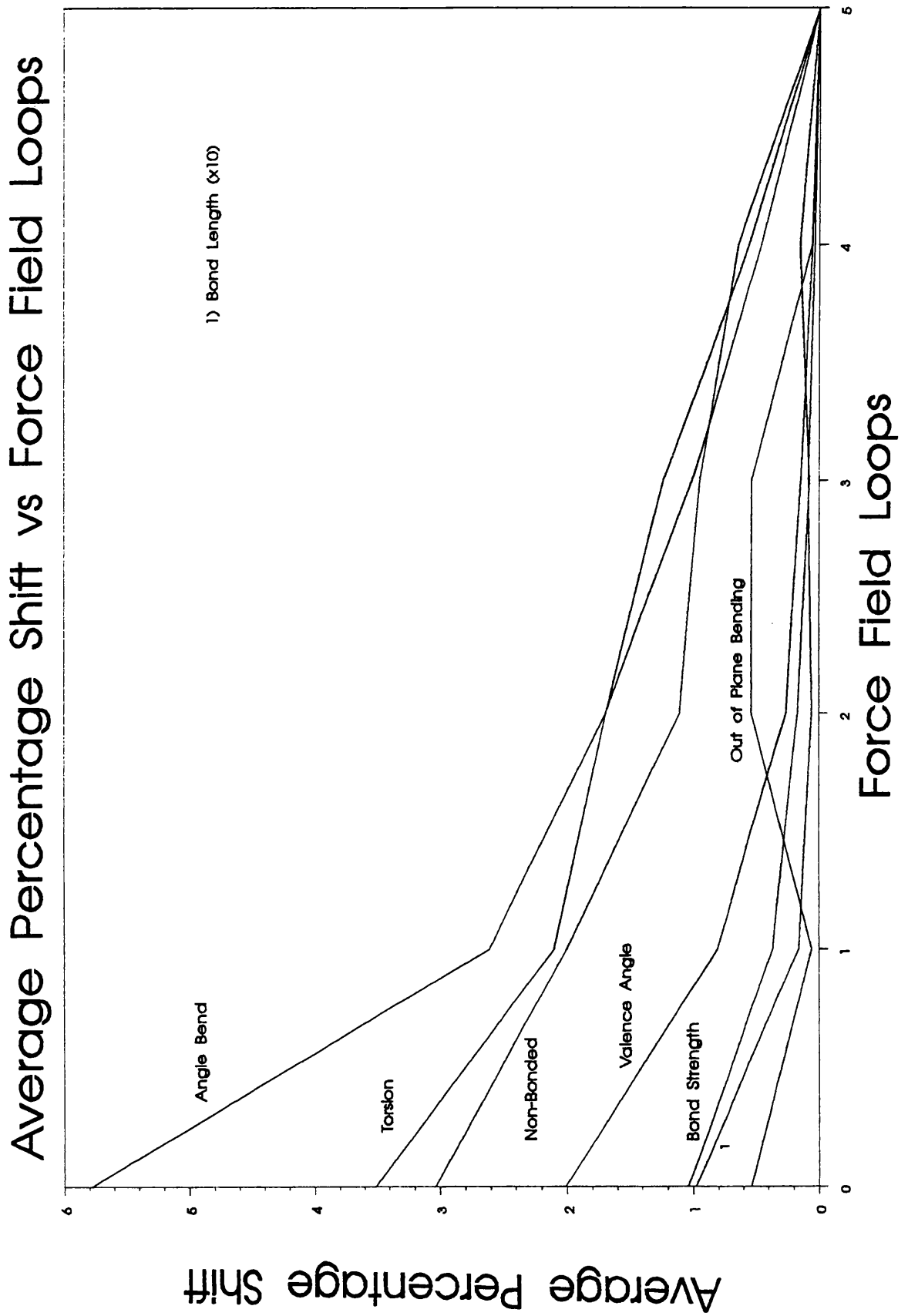


Figure 7.4 Typical Parameter Movements For Expanded Set



Another observation that can be made from figures 7.3 and 7.4 is the variation in the optimisation of different parameter types. This can best be explained by examining the quantitative effect of different parameter groups on the residual factor.

Consider, for example, the different parameters involved in the calculation of the bond angles centred at a given atom. The main parameters effecting the final minimised values of these angles will be the strain free valence angles and their associated angle bending force constants. It is reasonable to expect that the values for the valence angles are, largely, tied to a small range of values close to those determined theoretically from the appropriate orbital combinations. Thus it would be expected that any bond angle with a central C_{sp3} atom will be within a few degrees of 109.46°. This close tie between parameter and structure means that parameters such as the strain free bond angles and bond lengths tend to optimise far quicker than less specific parameters such as angle bend force constants and torsional barriers.

Many of the optimisations attempted to save time by optimising only one group of parameters for an initial set of force field loops and then optimising all parameters until termination. The following figures show the movements of the residual factors and the parameters for a typical example of this approach.

Figure 7.5 shows a plot of residual factor against iterations for a situation where all but the angle bending force constant parameter types have been fixed for the initial 20 iterations with all parameters optimisable for the remaining 9 iterations.

It can be seen that the curve of the initial section of the graph is similar to the runs shown previously with a rapid initial desent followed by an almost level section. Figure 7.6 shows the movement of the parameters being optimised in this section.

When all the parameters become optimisable there is a further drop in the residual factors as the other parameters adjust to the values of the parameters optimised in the first section.

Figure 7.7 shows, in the same manner as before, the movement of the parameters for the second stage of the optimisation.

It can be seen from these graphs that the effect of fixing some parameters initially is to produce effectively two optimisations, as the plots for each section show the same distinctive shape.

All these graphs show, then, that as the optimisation progresses the residual factors and the parameters tend towards some final values but that, in a similar way to exponential decay that only goes to zero at infinity, it never reaches these values.

Figure 7.5 Residual Factor Plot with Initially Only Angle Bending Parameters Optimising

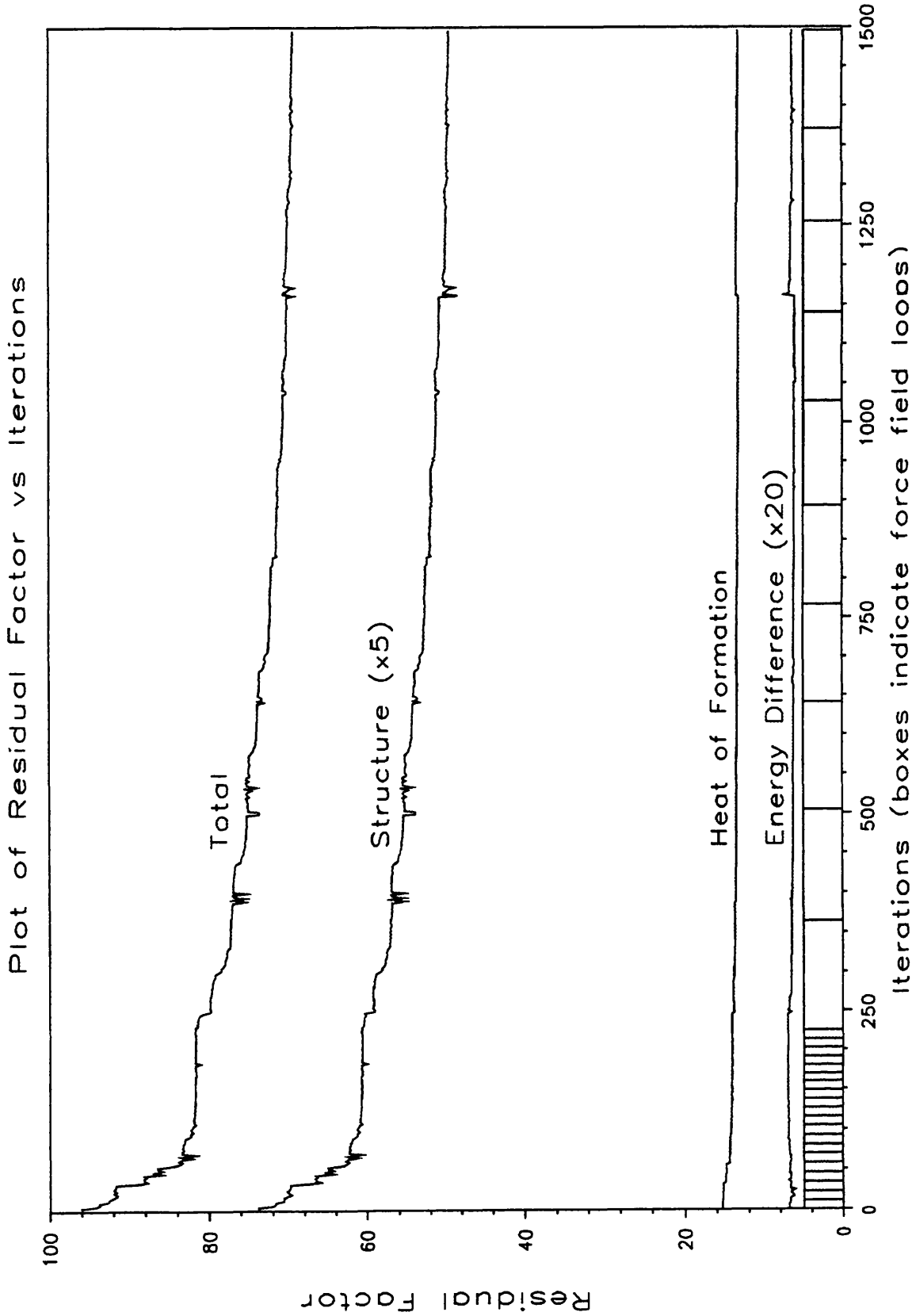


Figure 7.6 Movement of Angle Bending Parameters During First Section

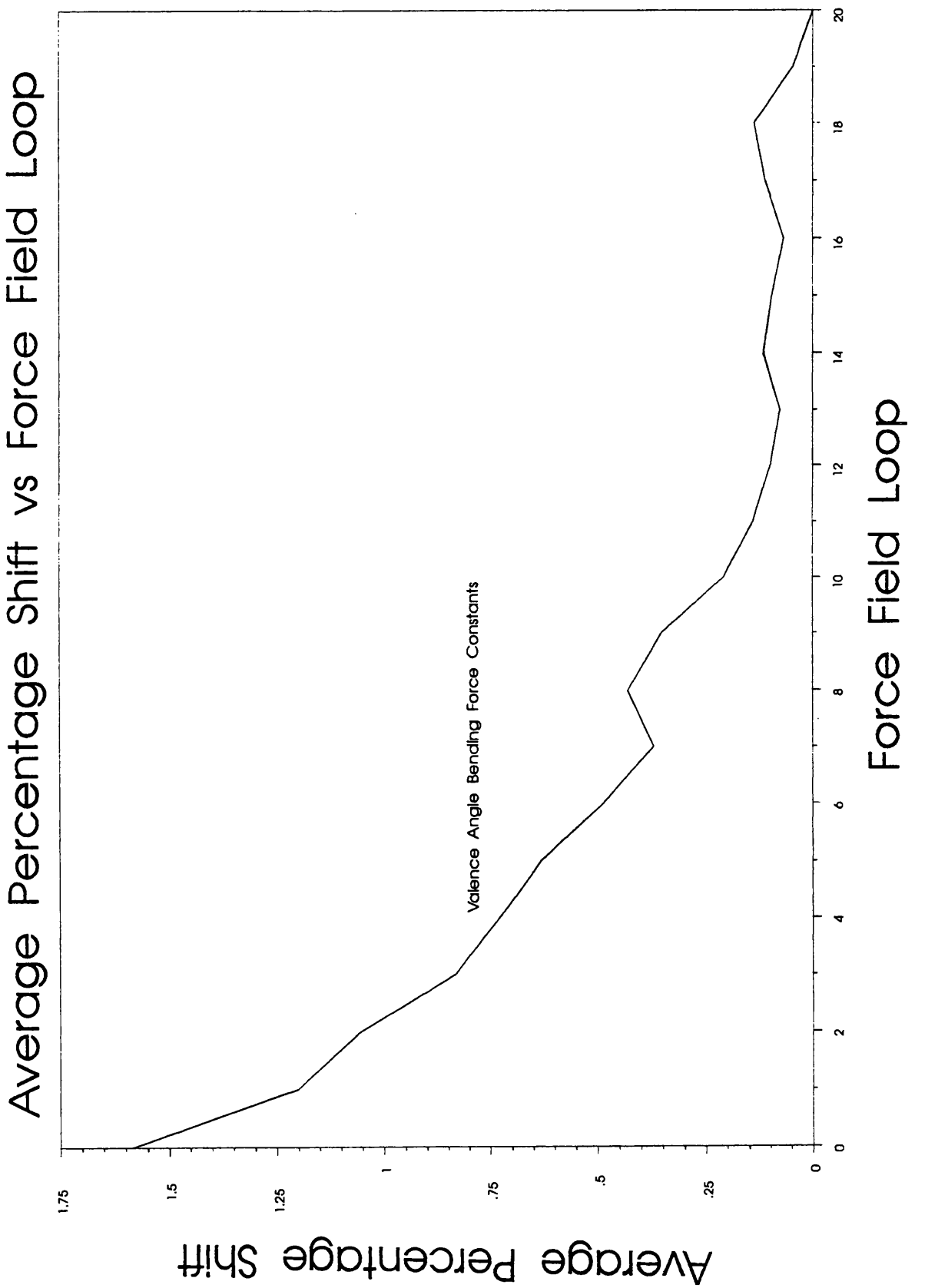
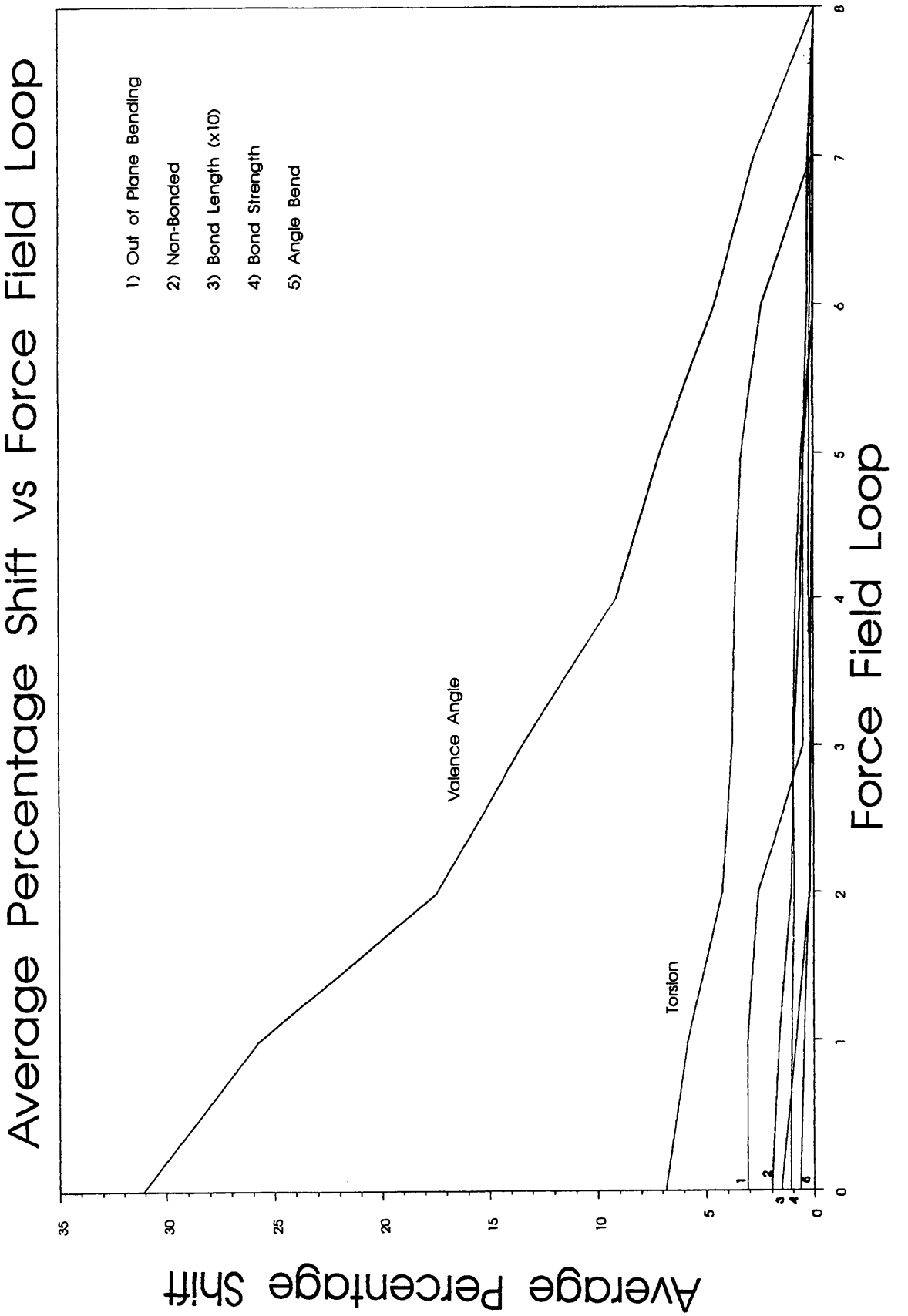


Figure 7.7 Movement of All Parameters During Second Section



7.2.2 Deficient Structure Sets

As well as the fact, as stated above, that the optimised parameters are effectively approximations to the 'optimum' values there are also the problems relating to deficiencies in the structure sets. These problems relate to the difference between the 'optimum' values for a given structure set and the 'optimum' values that would result from a more general structure set.

That is, it is possible to consider that for each force field set-up there are two optimum sets of parameters. The first is the 'optimum' values that would result in the lowest total residual factor for the current structure set and selected weightings. These shall be designated as the local optimum parameters. The second set of values can be considered to be those that would be produced by using the same weightings but with a theoretically complete structure set that would include all possible structural and thermochemical properties of all possible molecules. These shall be designated as the global optimum parameters.

The prime requisite of producing a general molecular mechanics force field must be that the local optimum parameters, produced by whatever method used, should be as near as possible to the global optimum parameters. It is only then that it is possible to have any trust in the results of molecular mechanics calculations performed on structures not included in the structure set used in the production of that force field. The problem is that the values for such a global structure set are, of course, not available for comparison. It is possible to examine the parameters produced from an optimisation and locate any values that are obviously unrealistic.

The values for the local optimum parameters will always be different to the global values as the local structure set is obviously a very small subset of the global case. It is hoped that the differences will be small but for some structure sets the local optimum parameters can end up being artefacts of the structure set deficiencies and totally unrelated to the global optimum values.

For example, given that the C=O bond length varies with the types of atoms connected to the Carbon (see section 7.4.1.1), then if a structure set had 90% of the structures with a Halogen atom shortening this bond then the 'optimum' bond length for this structure set would be shorter than that required for a more balanced structure set.

Alternatively, consider the problems involved in the determination of the equilibrium bond angle. When first considered it would appear that to reproduce the experimental results the strain free valence angle parameter should be close to the experimental values for a given situation. This is not necessarily true, especially in optimisation situations where there is a paucity of thermochemical data. For example, taking a simple case, if the central atom is a C_{sp3} and all the atoms connected are

identical, for example, all Hydrogens in methane, then it does not matter what the force field says each angle should be in isolation, it is obvious from symmetry that they will all end up being 109.47°. This is because the final angles after minimisation are at an equilibrium between all the forces involved. This reasoning can be extended to situations where the connected atoms are different for, roughly, as long as the ratio of the strain free angles are similar to the ratio of the angles determined experimentally, then the structural properties when minimised will be close to the experimental.

In these cases the main factor that stops the force field parameters from moving excessively away from their theoretical optimum values is the thermochemical data. That is, that though, for example, CF_4 will have the same structure independent of the F-C_{sp3}-F strain free valence angle parameters, it will not have the same steric energy. Thus the use of enough relevant thermochemical data should prevent this deviation from occurring.

It is quite likely, that, without the correcting effect of the thermochemical data, that such parameters will vary excessively from their optimum values. To demonstrate this point consider two cases, the first where the strain free valence angles are close to the physical values and the second where the ratios of the strain free valence angles are the same but are now scaled up by a factor of, say, 20%. If the same structure was then minimised with these two force fields it would be found that the internal co-ordinates would be relatively close. This can be demonstrated with the results of two runs of the CHEMMIN[2] molecular mechanics program, which uses the 'central times substitution factor' method of determining strain free angles. In the first investigation the base C_{sp2} angle was the default 122°, in the second case this was changed to an unrealistic 146°. Acetone was then minimised for both cases with the result that the largest difference in bond angles between the two cases was only 1.9°. Now if this small difference in the valence angle resulted in the calculated structure being closer to the experimental results then it can be seen that there is a structural based incentive for the strain free angle parameters to move to unrealistic values. The correcting effect of energy information can also be seen in this case as the relevant energies after minimisation were -3.65 Kcal/Mol and 15.31 Kcal/Mol. Although the creation of different group enthalpy increments for these two cases will reduce the effect of this large change in energy it should still give rise to an increase in the overall residual factor, thus providing an energy based incentive against the unreasonable deformation.

Another reason for having sufficient experimentally determined energy properties is that it is also possible for insufficiencies in the energy properties to

produce artefacts in a similar way to insufficient structural data. For example, when using a method of parameter determination that uses a large number of parameters, it is found that this has a greater effect on the heats of formation's residual factors than on those for the other properties. The most likely reason for this, especially for electron diffraction data, where not all structural properties are given, is that one or more of the parameters will be used in an interaction that has no directly related structural data and little effect on interactions where structural data is given. As such it is likely to result in unrealistic energy values, which, nonetheless, will, with the addition of the geometric enthalpy increments, produce a better residual factor than the true global parameter value would result in.

7.2.3 Torsion Angles

Of all the molecular properties being considered torsion angles provide one of the main problems. To understand the problem it is essential to distinguish what makes torsion angles different from the other structural properties.

This difference relates both to the ease of torsion angle deformation and that in most cases there will be several torsion angle values that will result in a local energy minimum.

The result of this is that torsion angles are particularly unstable during optimisation and several optimisation runs have run into difficulty because one or more of the torsion angles have moved well away from their experimental values. This is especially true when starting force field parameters are far from their 'optimum' values. In these cases torsion angles can move completely away from the experimental values early on in the optimisation, thus moving into different minima. The large residual factors produced have a detrimental effect on the overall optimisation. Because of this it is possible to say that a low standard deviation for the torsion angle residual factors indicated a good force field but that the opposite it not necessarily true.

In an attempt to counteract this problem the program was, towards the end of the project, adapted so that an artificial addition to the energy was made when a torsion angle moved outside 5 times its experimental error. This has the result of effectively restricting the torsion angles to these bounds. The worst possible scenario in this case would be that many of the torsion angles would finish up at, or just outside, these limits. This would result in higher energies from the artificial addition, adding artefacts to the energy calculations and that the torsion angles would be unrealistic compared to those that would be expected if the structure had been minimised without these limits.

It has been found that in most optimisations where this limit is in effect a few

of the torsion angles do end up at the range of this limit. This is considered acceptable as the effects mentioned above are small and incorrect torsion angles can affect all the other properties calculated for the minimised structures.

7.2.4 Residual factor Weightings

There are several points that need to be taken into consideration when trying to decide the weights to be applied to the residual factors. Though, as shall be discussed below, there are no firm rules, so the final choice of weightings will always tend to come down to personal opinion.

7.2.4.1 Legitimacy of Comparison, Errors and Reliability

When comparing calculated and experimental properties the main point to be considered is what situation does each case represent. If the cases are different then it is necessary to consider how comparable the cases are as a guide to the degree of weighting that should be used.

Structural Properties

Electron diffraction gives measures of the separation between all the atoms in molecule. A structure that reproduces this arrangement is then found. Because electron diffraction, like x-ray diffraction, takes place over a period of time the atomic positions can be considered as time averaged.

Simple molecular mechanics, i.e. not molecular dynamics, does not take into account any thermal properties such as vibration or rotation. Because of this molecular mechanics is often considered to produce the structures that would result if all thermal motion was removed from the molecules under investigation.

This, though, is not quite true as the equations used relate to this situation but the parameters used are derived using experimental results where there is thermal motion. Thus if the structure of a compound was determined at room temperature and at close to 0K, and these compared to the same structure determined using a molecular mechanics force field, then it would be expected that the modelled structure should be closer to the determined structure at room temperature, because this is the situation the force field has been optimised to attempt to reproduce.

This gives rise to interesting questions about the veracity of determining a force field in this way and then proceeding to use the same force field in molecular dynamics simulations where thermal motiona are added. In the future it may well be possible to produce a force field matching the results of averaged molecular dynamics studies and experiment and it would be intriguing to see what a comparison of the resultant force field with those currently determined using stationary structures would

reveal.

In the current discussion the main concern is the equivalence, or otherwise, of the structural properties between the experimental and molecular mechanics results. The main problem in this area comes, as stated above, from the fact that the experimental results are averaged over time, whereas molecular mechanics results in a single minimum energy 'snap shot'. In many cases, especially for small molecules, the global minimum will be significantly lower in energy than any other. In these cases the atoms will vibrate about their global minimum positions and the correlation between the experimental and the calculated results will be good. The difficulty comes when the global energy minimum is not low enough to contain the thermal motion of the atoms. An example of this problem is found with very flexible molecules such as proteins. When molecular dynamics simulations are done on proteins with x-ray co-ordinates as a starting point it has been found that initial relaxation of the protein is required to remove those bad interatomic interactions resulting from the time averaging of atomic co-ordinates[3].

Of all the structural property types the greatest problem in this case is with the torsion angles due to their previously stated ease of deformation. This problem is to some extent taken into account when using the values and errors from electron diffraction studies as the errors reported give a good indication if a problem with this property is likely to occur. For example, if a reported structure gives two torsion angles, 10.0°±3.0° and 60.0°±8.0°, then it can usually be assumed that the second torsion angle is easier to deform. In other words the structural properties likely to suffer most from the difference between the time averaged experimental observations and the 'snap shot' calculated results will tend to have the largest experimental errors. Because these errors are used to scale the individual residual factors before they are added to the total the effect of these differences is minimised.

So, in consideration, the situations being compared for the structural properties are not identical, but they are relatively close and the experimental errors associated with each structural property are likely to reflect the likely difficulty in each case.

Energy Difference Between Isomers

As discussed earlier in section 5.5.4.4, the calculated difference relates to the difference in the steric energies. For the experimental data the difference is, in fact, the difference in the enthalpies. As the enthalpies are not related to the entropy of the molecules these values are more or less directly comparable.

One problem that did arise is that the difference in the energy between isomers seems to vary depending on the method used to calculate the difference[4,5]. Most of

the experimental data for the conformers comes from electron diffraction studies, where the difference in the energy is calculated using a Boltzman factor from the temperature and the relative fractions of the various conformers. These values, if available, are used in an attempt to keep the various situations as consistent as possible.

Thus, again the quantities being compared are not totally identical, but again it is beneficial to continue with the comparison and take the difference into consideration when assigning weightings.

Heat of Formation

For the heat of formation there are two points to be taken into consideration. The first of these is similar to the energy difference above in that it is necessary to compare the type of measurement of heat of formation against the type of energy measurement produced from the molecular mechanics calculations.

The second point relates to the geometric enthalpy increments. The addition of these to convert the steric energy to the heat of formation is an empirical approach and as with other such methods, such as molecular mechanics, it will not be exact and so will give rise to errors of its own. The result of this is that the calculated values for the heats of formation can be considered to be less accurate than the calculated energy differences and structural properties.

7.2.4.2 Program Aims

As well as the more easily defined arguments involving errors and property comparability another element that needs to be taken into consideration when deciding on property weightings is the final aims of the force field.

By this, what is meant is, what are the priorities associated with the reproduction of the various molecular properties?

In the ideal case all three residual factors would be optimised to give perfect results, i.e. a zero value for each case. In the real world it is obvious that, for the structure residual factor at least, this situation is likely to be impossible to produce and so it is necessary to determine an order of priorities for the residual factors. For example, if the energy differences were too highly biased then it is likely that the final structures will be structurally deformed and similarly if the structure values are too highly weighted then it is likely that the energy difference residual factor will be unreasonably high.

Obviously a compromise position is required so that the result is a general force field that produces reasonable values for all possible properties.

7.2.4.3 Residual Factor Comparability

Residual factor comparability refers to how comparable are the residual factors before the application of weights. For example, how can a structure residual factor be compared with an energy residual factor? The nearest that can be done is to try to make sure that the values relate to an average property. That is, that the overall structure residual factor is the RMS of all the structure residual factors and so on for the other property types. This has been done in OPTPARAM for all the property types so in this case the 'playing field' is as level as it can be.

7.2.4.4 Range of Residual Factor Movement

When an optimisation takes place, for example the optimisations shown in figures 7.1 and 7.2, it is obvious that the range covered by each of the residual factor totals will be different. Generally what is found is that the structure residual factor tends to have a greater range. The problem here is that, for example, a drop in the structure residual factor from 20.0 to 15.0 at the expense of a rise in the energy difference residual factor from 1.0 to 3.0 will, with equal weighting, produce a lower total residual factor. This hides the fact that a 25% reduction in the structural residual factor will have taken place at the expense of a tripling of the energy difference residual factor.

7.2.4.5 Weighting Conclusion

It can be seen that there are many points to be taken into consideration when deciding the weighting of the residual factors. Because of the unquantifiable nature of several of these there can be no definitive rules to follow and in all cases it comes down to a personal decision taking all the above ideas into account. In the optimisations the weightings used were as follows:

Alkene Structure Set
Structure 1
Conformer 1
Heat of Formation 1
Expanded Structure Sets
Structure 5
Conformer 20
Heat of Formation 1

For the Alkene structure set there are only three conformer energy difference cases used so it was decided that these should not be significantly biased. For the heat of formation the use of group enthalpy increments gives only a small total

residual factor, typically 0.12, compared to a typical structural total of 0.6. So a one to one weighting in this case is unlikely to prefer the heats of formation over the structural properties.

For the Expanded structure set the structural residual factor is typically 10.0 as compared to an energy difference residual factor of 0.30. Because of this the selected weight for the energy difference is greater than that for the structural residual factor to make sure that, as mentioned above, the energy difference residual factor is not overwhelmed. For the heat of formation the general range of the residual factor depends much more on the number of parameters and can range from 1.0 up to 20.0. Remembering that extra errors are raised by the use of the group enthalpy increments and because of the possible large size of this residual factor it was decided that the weighting for the heat of formation residual factor should be less than that for the other residual factors.

7.2.5 Program Idiosyncrasies

When OPTPARAM was run the course of the optimisation occasionally took unexpected directions. In many of these cases these problems were traced to bugs in the code and corrected, but some others were either results of, presumably, bugs that were never found or an unexpected result of the deliberate programming strategy.

These problems affected the course of the optimisation in several ways. Either there was a large jump in the residual factor, from which the optimisation never recovered, or they caused slight blips in the overall residual factor. Considerable effort was expended trying to find the cause of these problems, but, due mainly to the long durations, up to several days, before these effects revealed themselves, not all the causes were tracked down.

The two main theories to explain these problems are as follows:

1) That the alteration in the force field distorts the energy surface to the point where the local minima is so shallow that there is little to drive the atom away.
2) Some error in the program causes atoms to be moved by distances greater than those normally allowable. This would then result in the co-ordinates either returning slowly to close to their original position or moving to a new conformation.

The difficulties in investigating these problems are that a large amount of information not only needed to be saved after every iteration but that this information would have to be recovered from the transputer network.

When an optimisation did contain a large jump in the total residual factor that run was terminated and the results not used in any further analysis.

7.2.6 Effects of Preliminary Consideration on Results Analysis

At this point it is useful to summarise the information given above and from this draw some deductions about the veracity of comparing the optimisation results.

1) Considering the points given in 7.2.4 a set of weighting values is decided.
2) These together with the chosen force field set-up define what will be called the global state. This information and the global structure set will have an equivalent set of global optimum parameters.
3) A structure set is defined and chosen. This information, with the force field set-up and the residual factor weightings, would, given complete optimisation, produce the local optimum parameters. As stated above, these local optimum parameters could contain values considered to be too much of an artifact of the structure set used to be generally applicable.
4) Optimisation of the initial approximate parameters brings the parameter values closer to the local optimum values. At program termination the difference between the current parameter values and the local optimum values is, of course, unknown.

Because of all the points in the above list it can be seen that it is only possible to make deductions about the results on a generalised basis. That is to say, for example, that it would be unrealistic to state, categorically, the best method of determining the parameters for a particular interaction type if the total residual factors for the different methods are close. Instead the distribution and statistics of the residual factors and the number of parameters that would be required for a full force field are examined and the option that appears to give the best balance is chosen.

7.3 Structures Used

7.3.1 Alkene Structure Set

The final data set of structures for the alkene force field consisted of 50 structures. Of these, 21 contributed to the structural residual factor, 38 to the heat of formation residual factor and 6 to the residual factor for energy differences between isomers.

The following is a list of the structures used and indicates for each case which of the residual factors it is involved in the calculation of. ST stands for structural residual factors, ED for energy difference between isomers and HF for heat of formation[6].

Structure	ST	ED	HF
Gauche butane		X[7]	
Anti butane		X[7]	
Cis but-2-ene	X[8]	X[9]	X
Trans but-2-ene	X[8]	X[9]	X
Boat chair bicyclo[3:3:1]nonane		X[10]	
Twin chair bicyclo[3:3:1]nonane		X[10]	
Adamantane	X[11]		X
Bicyclo[2:2:2]oct-2-ene	X[12]		
Bicyclo[2:2]octa-2,5-diene	X[12]		
Cis but-1-ene	X[13]		
Cyclodeca-1,6-diene	X[14]		
Chair chair bicyclo[3:3:2]decane			X
Cis hex-2-ene			X
Cyclohexa-1,4-diene	X[15]		X
Cyclohexene	X[16]		X
Cycloheptene			X
Cyclopentene	X[17]		X
2,3-Dimethyl but-1-ene			X
2,3-Dimethyl but-2-ene	X[18]		X
4,4-Dimethyl pent-1-ene			X
Ethylene	X[19]		X
Hexa-1,5-diene			X
Isobutene	X[20]		X
1-Methyl adamantane			X
1-Methyl cyclohexene			X
Manxane	X[21]		
2-Methyl skew but-1-ene	X[22]		
1-Methyl cyclopentene			X
1-Methyl diadamantane			X
2-Methyl but-2-ene			X
2-Methyl pent-2-ene			X
3-Methyl but-1-ene			X
4-Methyl cis pent-2-ene			X
4-Methyl trans pent-2-ene			X
2-Methyl but-1-ene			X
Norbornadiene	X[23]		X

Structure	ST	ED	HF
Norbornane			X
Pent-1-ene			X
Pin-2-ene	X[24]		
Propene	X[25]		X
Penta-1,4-diene			X
Peri trihydro quinacene			X
5-Skew cis pent-2-ene			X
5-Skew trans pent-2-ene			X
Skew but-1-ene	X[26]		
Trans trans trans cyclododeca-1,5,9-triene			X
Cyclohexane	X[27]		X
Trans hex-2-ene	X[28]		X
1,3,5,7-Tetramethyl adamantane			X
2,4,4-Trimethyl pent-1-ene			X

7.3.2 Expanded Structure Set

As stated in the introduction it was found that the use of a simple alkene force field involved too narrow a data set to fully check some of the force field options. In particular it was thought that the techniques to determine the bond angles and the torsion barriers needed further investigation with a more varied structure set. The extra atom types chosen for this structure set are Fluorine, Chlorine, Bromine, and the double bonded Oxygen O_{sp2}.

There are two reasons for choosing these atom types. The first is that, as they are all terminal atom types, it means that the differences between situations for the valence angle and torsion twist cases are greater in proportion to the number of parameters than if alternate non-terminal atom types had been chosen. For example, consider the different methods for deciding on the torsion barrier parameters. It can be seen, because the only additions are terminal atom types, that there will be the same number of parameters required if only the central two atom types are considered. If, however, the method using all four atom types is used then there is a large increase in the number of possible cases. Thus the ratio of parameter numbers for the two cases is considerably higher in the expanded data set.

The other reason for choosing these atom types is simply that structures with these atom types constitute a relatively large proportion of all the structures

determined by electron diffraction and so there is a significant database of relevant structures that can be drawn on.

During the investigations with these atom types two structure sets were used, with the second set being an extension of the first. In the following lists of the structures the first section relates to structures used in both sets and the second section relates to those structures added in the second set. Where no reference is given for the energy difference it implies that the relevant thermochemical data was obtained from the same reference as that relating to the structural data.

First Expanded Structure Set

Structure	ST	ED	HF
1-Bromo 2-chloro 1,1,2-trifluoro ethane			X
1,1-Chloro ethylene	X[29]		X
1,2-Dibromo propane	X[30]		X
1,2-Dibromo 1-chloro 1,2,2-trifluoro ethane			X
1,2-Dichloro propane	X[30]		X
1,3,5,7-Tetramethyl adamantane			X
1-Methyl adamantane			X
1-Methyl cyclohexane			X
1-Methyl cyclopentene			X
1-Methyl diamantane			X
2,3-Butanedione	X[31]		X
2,3-Dimethyl but-2-ene	X[18]		X
2,3-Dimethyl but-1-ene			X
2,4,4-Trimethyl pent-1-ene			X
2-Chloro cyclo hexanone	X[32]		
2-Methyl but-1-ene			X
2-Methyl but-2-ene			X
2-Methyl pent-2-ene			X
2-Methyl skew but-1-ene	X[22]		
3-Methyl 2-butenoyl chloride	X[33]		
3-Methyl but-1-ene			X
4,4-Dimethyl pent-1-ene			X
4-Bromo but-1-ene	X[34]		
4-Methyl cis pent-2-ene			X
4-Methyl trans pent-2-ene			X
5-Skew cis pent-2-ene			X

Structure	ST	ED	HF
5-Skew trans pent-1-ene			X
Acetaldehyde	X[35]		X
Acetone	X[36]		X
Acrolein	X[37]		
Adamantane	X[11]		X
Anti 1,1,2,2-tetrafluoro ethane*	X[38]	X	
Anti 1-bromo propane		X[39]	
Anti 1-chloro 2-bromo ethane		X[40]	
Anti trans 2-butenoyl chloride	X[33]	X	
Anti 1,2-difluoro ethane	X[41]	X	
Anti anti fumaryl chloride	X[42]	X	
Anti butane		X[7]	
Anti syn fumaryl chloride	X[42]	X	
Bicyclo[2.2.2] oct-2-ene	X[12]		
Bicylo[2.2] octa-2,5-diene	X[12]		
Boat chair bicyclo[3.3.1] nonane		X[10]	
Bromo chloro methane	X[43]		
Bromo dichloro methane	X[43]		
Carbon tetrachloride			X
Carbon tetrafluoride			X
Carbonic dichloride*	X[44]		X
Carbonic difluoride*	X[45]		X
Carbonyl bromide	X[46]		X
Chair chair bicyclo[3.3.2] decane			X
Chloro trifluoro methane			X
Cis but-2-ene	X[8]	X[9]	X
Cis 1-chloro 3-fluoro propene	X[47]		
Cis but-1-ene	X[13]		
Cis hex-2-ene			X
Cyclo heptanone	X[48]		X
Cyclo hexanone	X[49]		X
Cyclo deca-1,6-diene	X[14]		
Cyclo heptene			X
Cyclo hexa-1,4-diene	X[15]		X

Structure	ST	ED	HF
Cyclo hexane	X[27]		X
Cyclo hexene	X[16]		X
Cyclo pentanone	X[50]		X
Cyclo pentene	X[17]		X
Dibromo chloro methane	X[43]		
Dichloro difluoro methane			X
Ethylene	X[19]		X
Formaldehyde	X[35]		X
Gauche 1,1,2,2-tetrafluoro ethane*	X[38]	X	
Gauche 1-bromo propane		X[39]	
Gauche 1-chloro 2-bromo ethane		X[40]	
Gauche 2,3-dibromo propene	X[51]		
Gauche 1,2-difluoro ethane		X[41]	
Gauche butane		X[7]	
Hexa-1,5-diene			X
Isobutene	X[20]		X
Isopropyl carboxaldehyde	X[52]		
Manxane	X[21]		
Norbornadiene	X[23]		X
Norbornane			X
Pent-1-ene			X
Pentafluoro ethane	X[53]		
Penta-1,4-diene			X
Peri trihydro quinacene			X
Pin-2-ene	X[24]		
Propanal skew	X[54]		
Propanal syn	X[54]		
Propene	X[25]		X
Skew but-1-ene	X[26]		
Syn trans 2-butenoyl chloride	X[33]	X	
Tetra methyl p-benzoquinone	X[55]		
Trans but-2-ene	X[8]	X[7]	X
Trans trans trans cyclododeca-1,5,9-triene	X[56]		X
Trans hex-2-ene			X

Structure	ST	ED	HF
Trichloro fluoro methane	X[57]		X
Twin-chair bicyclo[3.3.1] nonane		X[10]	

Extra Structures in Second Expanded Structure Set

Structure	ST	ED	HF
1,1-Difluoro ethene	X[58]		X
1,1,1,2-Tetrafluoro ethane	X[59]		
1,1,1-Trifluoro ethane	X[60]		X
1,1,2-Trichloro 3,3-difluoro prop-1-ene	X[61]		
2-Bromo propene	X[62]		
2-Chloro propene	X[62]		X
3-Bromo propene	X[63]		X
3-Bromo 2-methyl prop-1-ene	X[64]		
Acetyl fluoride	X[65]		X
Anti 1,2-dibromo ethane	X[66]	X	
Anti 1,2-dichloro ethane	X[67]	X	
Anti 2,3-dichloro buta-1,3-diene	X[68]	X	
Anti 2,3-dichloro prop-1-ene	X[69]	X	
Anti 2-bromo 3-chloro propane	X[70]	X	
Anti 2-bromo 3-chloro 2-methyl propane	X[71]		
Bromo acetyl bromide (gauche)	X[72]		
Bromo acetyl chloride (gauche)	X[72]		
Bromo chloro fluoro methane	X[73]		
Chloro acetyl chloride (anti)	X[74]	X	
Chloro acetyl chloride (gauche)	X[74]	X	
Chloro acetaldehyde	X[75]		
Cis 1-chloro buta-1,3-diene	X[76]		
Cis 1,2-difluoro ethene	X[77]		
Dichloro acetyl chloride (gauche)	X[78]		
Formyl fluoride	X[79]		
Gauche 1,2-dibromo ethane	X[66]	X	
Gauche 1,2-dichloro ethane	X[67]	X	

Structure	ST	ED	HF
Gauche 2,3-dichloro buta-1,3-diene	X[68]	X	
Gauche 2,3-dichloro prop-1-ene	X[69]	X	
Gauche 2-bromo 3-chloro 2-methyl propane	X[70]		
Gauche 2-bromo 3-chloro propene	X[70]	X	
Hexafluoro acetone	X[80]		
Hexafluoro propene	X[81]		
Isoprene	X[82]		X
Monobromo ethane			X
Monochloro ethane			X
Monofluoro ethane	X[53]		
Perfluoro norbornadiene	X[83]		
Propionyl chloride	X[84]		
Skew 3-methyl 2-butanone	X[85]		X
Syn syn fumaryl chloride		X[42]	
Tetra fluoro p-benzoquinone	X[55]		
Trans 1-chloro buta-1,3-diene	X[76]		
Trifluoro acetyl fluoride	X[86]		
Trifluoro ethene	X[87]		X
Vinyl bromide	X[88]		X
Vinyl chloride	X[89]		X
Vinyl fluoride	X[90]		X

* Molecules for which energy data was used in the first set but where structure data was only added for the second expanded structure set.

7.4 Force Field Development

In the following sections each parameter section will be examined in turn. Each section is divided into discussions of the Alkene and the Expanded structure set.

Three types of graphs will be included in these sections. The first type simply shows the sum of the residual factors resulting from the optimisations. The data required for these graphs is obtained from the short summary recorded with each set of parameters. As this information is available for all optimisation runs such graphs can be constructed in all cases. As these graphs do not contain much information they are used only when little other data is available.

The second type of graph shows a distribution of the residual factors. That is,

for each property the error band that that property fits into is calculated. The number of properties fitting into each band for all the property types are then totalled. These numbers are both given numerically and displayed graphically as histograms. For the Alkene case a single graph is displayed showing the residual factor for the Bond lengths, Valence angles, Torsion angles, Energy Differences and Heats of Formations. For the Expanded structure sets there are more experimentally determined properties and so the graphs for these sections will be given separately. The data for this section was obtained from two file types, the saved COMPARE and RESULTS files. The best accuracy for the structural properties is obtained from the RESULTS file which is only present for later OPTPARAM runs. The RESULTS file does not contain the information relating to energy differences and heats of formation and so that has to be extracted from the COMPARE file. When a RESULTS file is not available it is also possible to extract the structural residual factor data from the COMPARE file. A typical COMPARE file is shown in appendix B.

In the third type of graph the distribution of the residual factors displayed in previous graph types are used to calculate the standard deviations for the residual factors of each property type. These are then given both numerically and as two histograms. The first histogram groups the results with the method of parameter determination and the second groups the results by property type for a more direct comparison. Both show the number of parameters used in each case for the interaction under consideration at the time.

7.4.1 Bond Stretching

Bond stretching involves two parameters for each interaction, the bond length and its associated bond stretching force constant.

7.4.1.1 Bond Length

The bond length parameters are singular amongst molecular mechanics parameters in that there is, in general, only one way that they are defined in molecular mechanics programs. That is, there is a different bond length between each combination of atom types, depending only on the bond multiplicity.

One exception to this is that different values can be used in small rings such as cyclobutane, but in most cases the atoms involved in these rings are generally considered to be different internal atom types. None of the structures used in any of the current optimisations involved any ring smaller than cyclopentane so it was considered that no special parameters, or atom types, would be used, though some force fields do define special parameters for rings up to this size[91].

This is not to say that the use of a single bond length is the ideal situation, especially when atom types with highly different electronegativities are used. The problem in this case is that it ignores the effect of other atoms connected to the atoms involved in bond. This shows up, for example, in examination of the $C_{sp2}=O_{sp2}$ bonds used in the optimisation. They show a large range of values, far greater than the general level of experimental error. The result of this is that OPTPARAM will produce a compromise value, resulting in high, i.e. bad, values for most of the relevant residual factors. There can, of course, be other interactions, mainly non-bonded interactions, that affect the bond length but in most cases there is little or no steric hindrance and so little to stop the bond length achieving the strain free value from the force field. This is especially true in terminal bonds such as the $C_{sp2}=O_{sp2}$ bond described above.

This variation in bond length with the type of substituents can be seen in figure 7.8 that shows the distribution of the experimental bond lengths of the $C_{sp2}=O_{sp2}$ bond. To produce this graph three sets of 0.001Å bins were set up, where each set of bins relates to the number of halogens connected to the carbon (0,1 or 2). Each bond was then examined to determine which bins fell within range of the experimental error. To each of these bins, for the set relating to the number of connected halogens, a number was added inversely proportional to the total number of bins. This number was given by the equation:

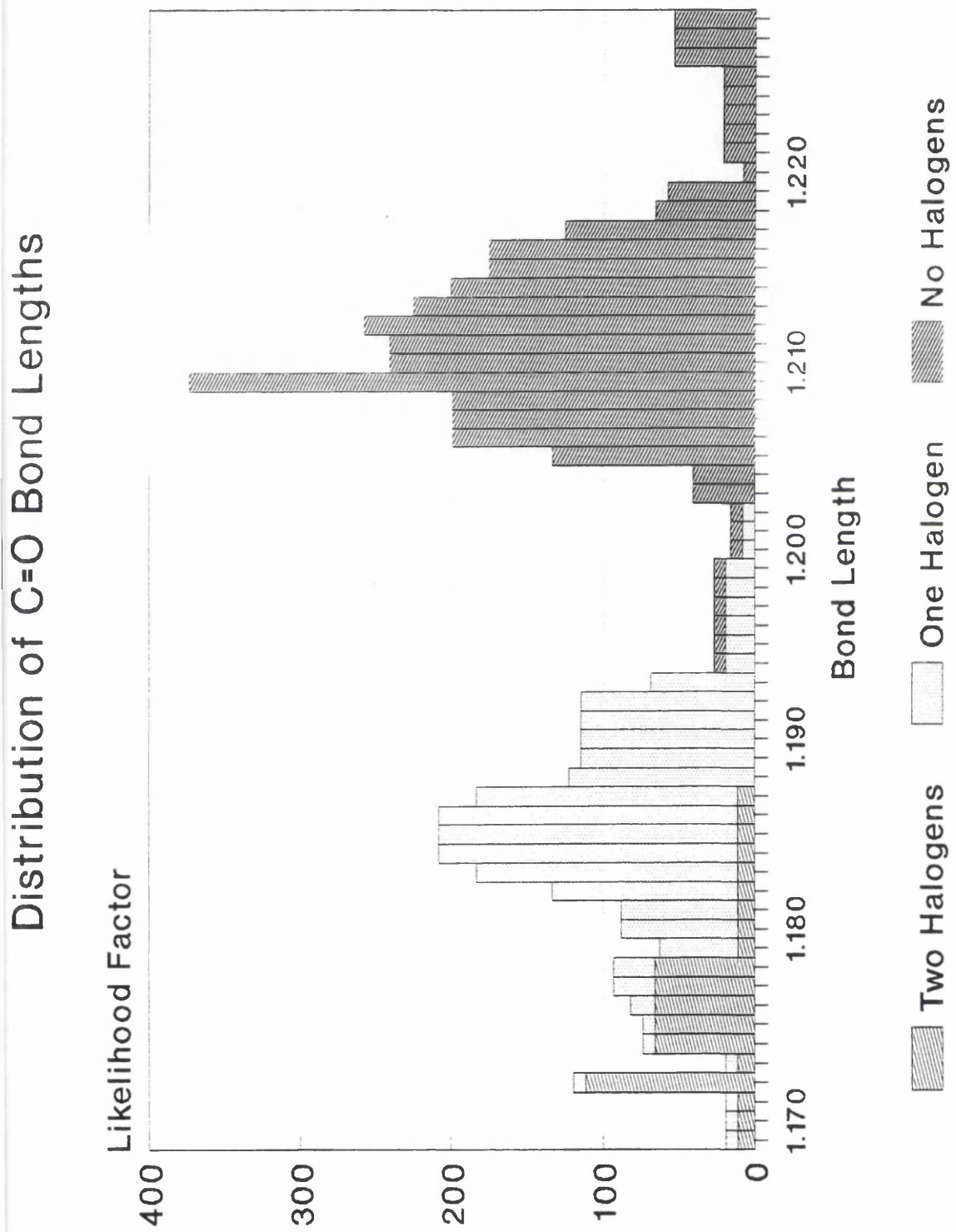
$$\frac{100}{\text{Number of Bins}}$$

Thus the same amount was added in total for each bond. The graph was then produced by totalling these three sets of bin results while still indicating the amount being contributed towards each bin from each set. The vertical axis is labelled likelihood rating as the number gives an indication of how likely it would be to find a bond where the experimental error included that distance.

It can be seen that the overall shape of the curve comes from the superposition of three distinct peaks, each relating to a different substitution situation. It can also be seen how the optimised bond length in this case is likely to be at a compromise position of about 1.2Å, a value that falls within the experiment error of only a few of the bond lengths. This is confirmed by the final optimisation, which as shall be seen later, produces a C=O bond length of 1.196Å.

It is possible to see from this that to get a more accurate representation of the bond lengths a more complicated situation than that usually used would be required. However, in determining the final structure and energy of a molecule the bond lengths are considered less important than both the valence and torsion angles,

Figure 7.8 Distribution of $C_{sp2}=O_{sp2}$ Bond Lengths



although they will, of course, affect the final values for these properties. Because of this it is possible to do realistic modelling, even if the bond lengths are not as accurate as they could be in all cases. A point proven by the fact that the vast majority of modelling done to date has been done using force fields with such singular bond lengths.

It should be possible to produce a function for bond lengths, similar to that discussed in chapter 4 for valence angles, which could be used to calculate bond lengths, taking into account the effect of the types of all associated atoms. However, it was thought that attention should be concentrated on the other, more important, interactions.

7.4.1.2 Bond Stretch

There are though, several possible ways of defining the bond stretch force constant parameters. The two methods examined in this study are the explicit determination of each possible value (full), i.e. each different bond length will have its own force constant, and the method as given in section 4.2 by which the bond stretch force constants are determined from the associated bond length using general equations involving a small number of empirical parameters (calculated).

Chapter 4 describes several other ways of determining the bond stretch force constant parameters. It is appropriate at this point to discuss why only two of the possible options were investigated.

The reason why the fixed bond length, i.e. infinite bond stretch force constants, situation was not investigated is that, though this approach would produce a reduction in the total number of parameters it would, as stated in section 4.2, require that internal rather than Cartesian co-ordinates be used during the molecular mechanics minimisation. Because of this a fixed bond length scheme was impossible to implement within OPTPARAM in its current form without completely rewriting large portions of the code. This effort was not considered worthwhile when the difference in the number of parameters between this and the 'calculated' case is so small compared to the size of the full force field.

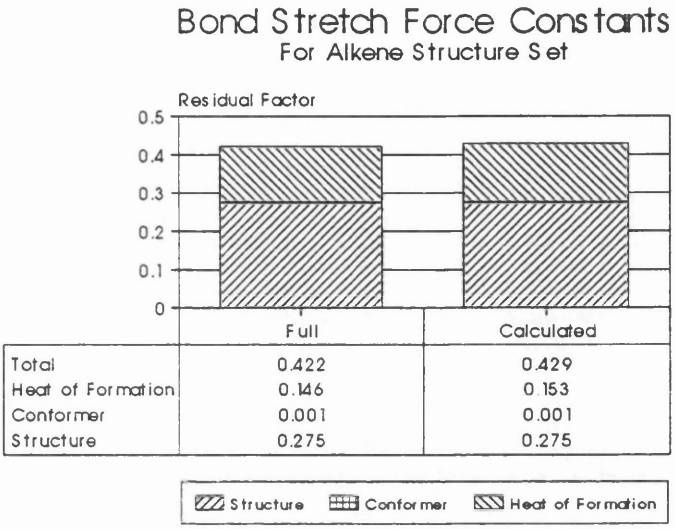
The same logic was applied to the situation using a single bond stretch force constant for conjugated and non-conjugated bonds, here the difference is only 3 parameters between this case and the 'calculated' case.

7.4.1.2.1 Alkene Structure Set

During the development of OPTPARAM several runs were made with identical force fields, except that in one case the bond stretch force constants were determined explicitly and in the other they were calculated. Figure 7.9 shows the

residual factor at the end of one of these pairs of runs.

Figure 7.9 Alkene Bond Stretch Force Constant Comparisons



It can be seen that the residual factors in these cases are almost identical. Processing of the relevant COMPARE files to produce the distribution of individual residual factors that also shows very similar values for both cases (figure 7.10).

Both of these graphs show that there is little or no difference between the 'full' and the 'calculated' bond stretch situations. So for the alkene situation at least, it can be said that the 'calculated' method of data reduction is a feasible alternative to the 'full' case. It has though to be remembered that this does not result in a significant decrease in the number of parameters that need optimisation as the number of parameters required for the both methods are roughly equal.

7.4.1.2.2 Expanded structure set

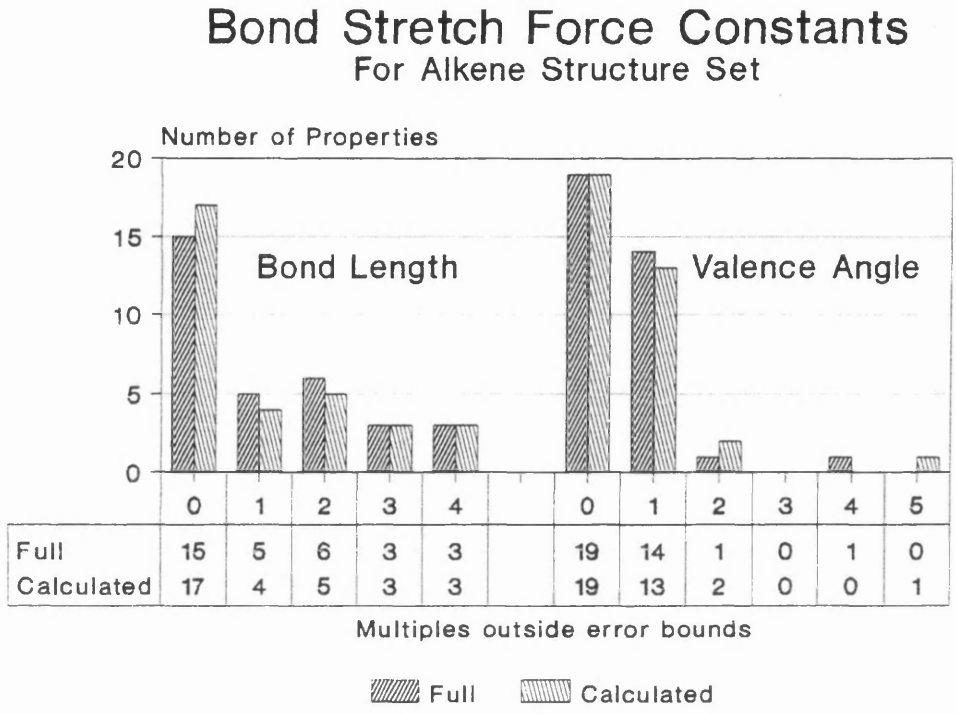
To further investigate the feasibility of using calculated bond stretch force constant parameters two runs were made with the first expanded structure set using the 'full' and the 'calculated' options.

Figure 7.11 shows a comparison of the bond lengths for these cases. Figure 7.12 shows a comparison of the other properties.

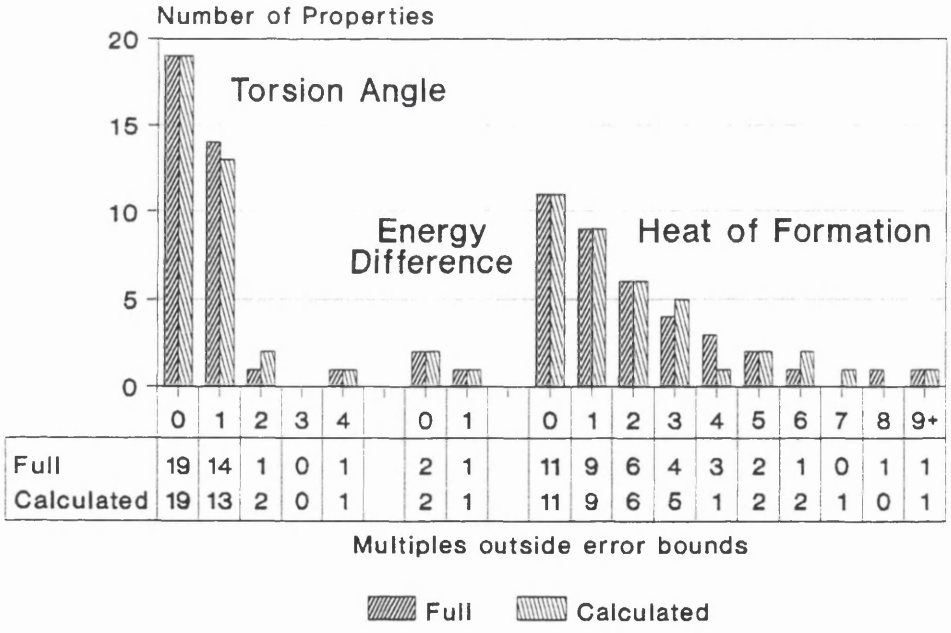
It can be seen in these graphs that there is very little difference between the results obtained for these two approaches and in this case the 'calculated' method uses 5 parameters, as in the alkene structure set, compared to a total of 13 bond stretch force constants for the fully defined option.

This closeness is also shown in table 7.1 which shows the average absolute difference between the experimental and calculated bond lengths compared with the average experimental error. That is, that the average difference between the calculated structural properties and the central experimental value is determined and this is

Figure 7.10 Residual Factors for Bond Stretch Force Constants, Alkene Study



	Bond Length					Valence Angle					
	0	1	2	3	4	0	1	2	3	4	5
Full	15	5	6	3	3	19	14	1	0	1	0
Calculated	17	4	5	3	3	19	13	2	0	0	1



	Torsion Angle					Energy Difference		Heat of Formation								
	0	1	2	3	4	0	1	0	1	2	3	4	5	6	7	8	9+
Full	19	14	1	0	1	2	1	11	9	6	4	3	2	1	0	1	1
Calculated	19	13	2	0	1	2	1	11	9	6	5	1	2	2	1	0	1

Figure 7.11 Distance Residual Factors for Expanded Structure Set

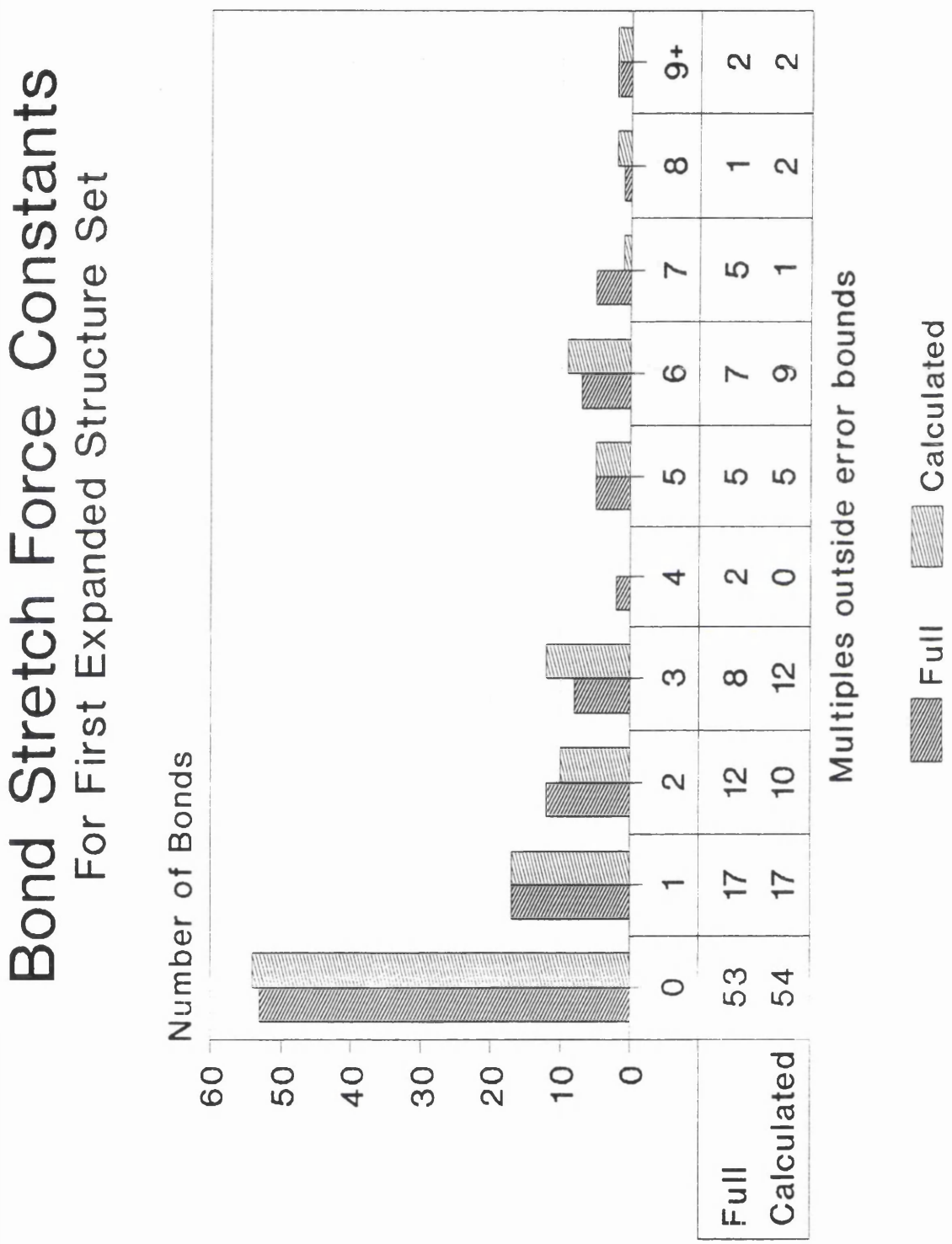
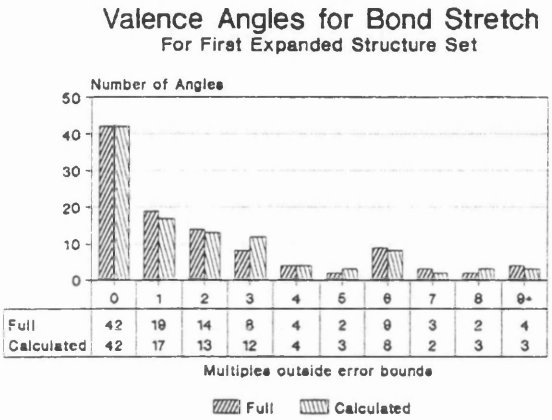
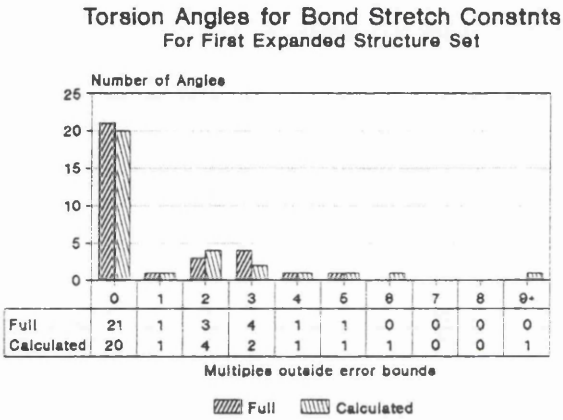
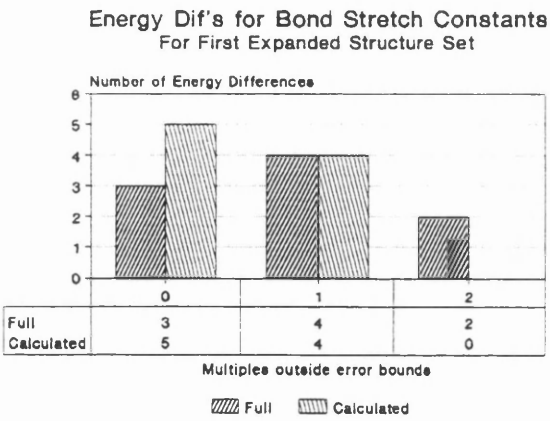


Figure 7.12 Other Residual Factors for Expanded Structure Set







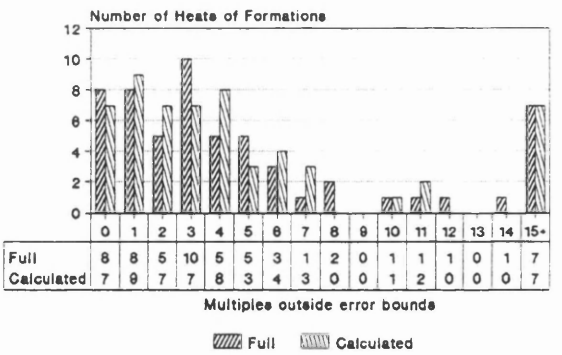
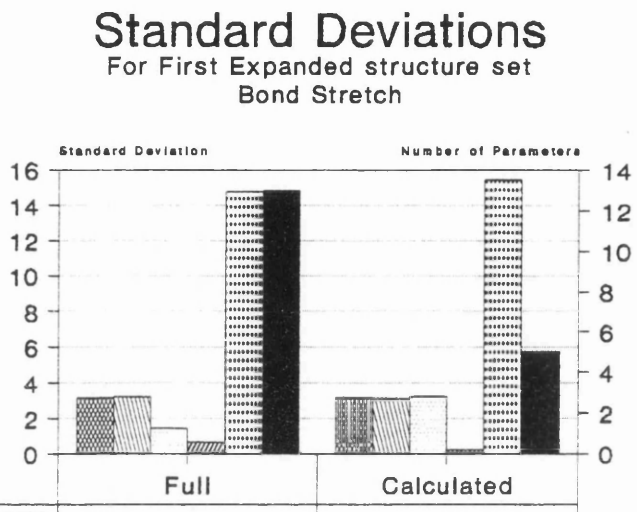
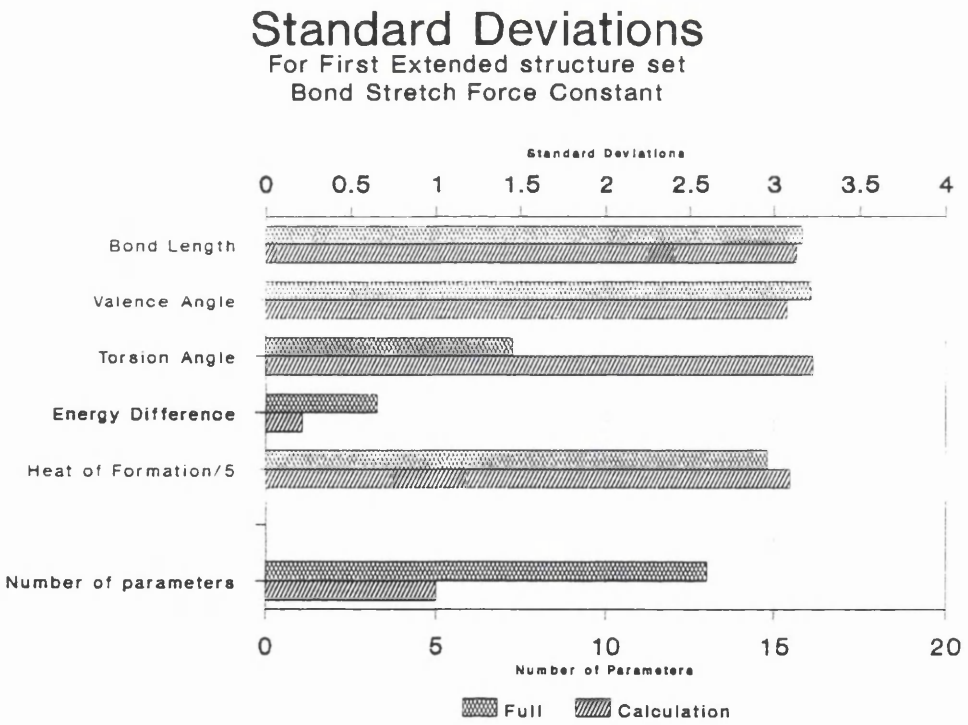


Figure 7.13 Standard Deviations for Bond Stretch Force Constant Options



Bond Lengths	3.164	3.13
Valence Angles	3.218	3.074
Torsion Angles	1.455	3.229
Energy Differences	0.654	0.217
Heat of Formation	14.788	15.464
Number of Parameters	13	5



compared with the average errors for these properties. Note that in both cases the average difference between the calculated and the experimental values compares very favourably with the average error in the experimental bond lengths.

Table 7.1 Comparison of Average Structural Property Differences with Errors

Force Field Type	Average Difference of Bond Length from Experimental
Full	0.0121Å
Calculated	0.0123Å
Average Error	0.0424Å

As a final confirmation of this closeness of the two cases figure 7.13 shows the standard deviation for the values plotted in figures 7.11 and 7.12. Considering the earlier discussion on the comparability of results it is not possible to say on these results alone which is the optimum method of parameter determination.

It is only on examining the actual values of the force constants, both 'full' and 'calculated', is it found that the values for these cases are, in some places, remarkably different.

Table 7.2 Comparison of Optimised and Standard Force Constants.

Bond	Full Force Constant	Calculated Force Constant	Standard Force Constant
C_{sp2}-H	316.9	313.1	444
C_{sp3}-H	319.8	304.9	438(1)
C_{sp2}=C_{sp2}	719.7	739.8	720(8)
C_{sp2}-C_{sp2}*	368.7	540.9	282(9)
C_{sp2}-C_{sp3}	422.3	519.6	------
C_{sp2}=O_{sp2}	699.4	1004.8	532(1)
C_{sp2}-F	514.1	612.2	------
C_{sp2}-Cl	248.7	308.0	310(13)
C_{sp2}-Br	187.2	232.9	------
C_{sp3}-C_{sp3}	404.3	487.3	376(2)
C_{sp3}-F	554.8	752.8	490(25)
C_{sp3}-Cl	183.5	312.5	306(7)
C_{sp3}-Br	139.5	226.1	225(7)

* Conjugated bond.

Table 7.2 gives a comparison between bond stretch force constants optimised normally and the resultant values produced from the 'calculated' option for the same bonds. As an extra comparison some sample values, where available, were taken from a standard reference source[92]. All values are in Kcal Mol^{-1} $Å^{-2}$.

It can be seen in most cases that the values produced using both approaches to parameter specification are reasonably close. There are though cases, such as C_{sp3}-Br, where there is a large difference between both the two approaches and the experimental values. The probable reason for this is, as stated before, that when a value is optimised and there is little data, especially thermochemical data, relating to appropriate molecules containing the given bond, then the optimisation can cause the value of the bond stretch force constant to become unrealistic to fit the small selection of experimental data used. This is especially a problem for bond lengths as the majority of molecules will not contain significantly deformed bonds so the number of molecules where any parameter will have a significant effect is small.

Because of these problems it can be expected that the bond stretching force constants, whichever way they are determined, will be difficult to optimise to produce chemically realistic values. One way to overcome this problem, not included in OPTPARAM because of the limited effect of the bond stretching force constant parameters, would be to include information about the infra red frequencies of the bonds in question. In the end it may not be possible to use the OPTPARAM program to produce the bond stretching force constants required, but that importing more realistic values from sources, such as other force fields, will have only a very small effect on the overall residual factor.

7.4.2 Angle Bending

As discussed in chapter 4 the determination of the angle bending parameters has the largest possible combination of parameter selection options. This is because each angle will require a strain free angle value and at least one bending force constant to determine the difficulty of the bond bending and that there is no requirement for the methods used to be connected in any way. That is, if the strain free valence angle is determined on a 'full with substitution factor' basis, there is no requirement that the angle bend force constant be determined in the same way.

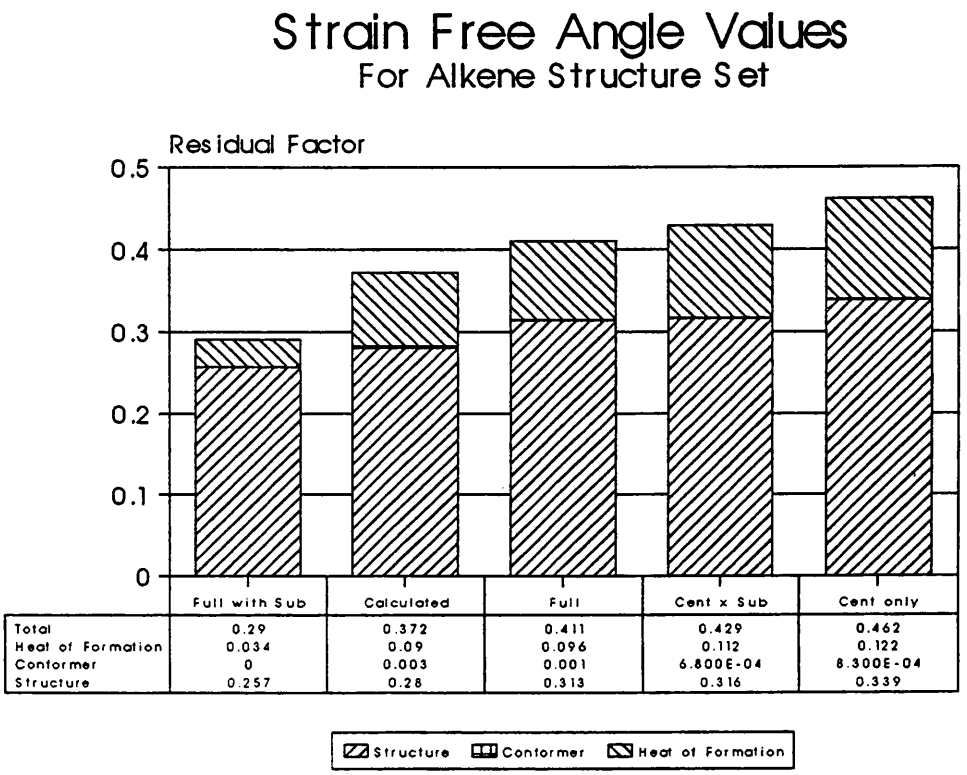
Each of the parameter groups in question shall be considered in turn, starting with the methods used to assign the various strain free bond angles. As before the results for the alkene investigation will be given and these followed by those relating to one or both of the Expanded structure sets.

7.4.2.1 Strain Free Valence Angle

7.4.2.1.1 Alkene Structure set

Figure 7.14 shows the residual factors from OPTPARAM runs on the alkene structure set where only the method of determining the strain free valence angle has been varied.

Figure 7.14 Total Residual Factors for Strain Free Angle Options, Alkene Study

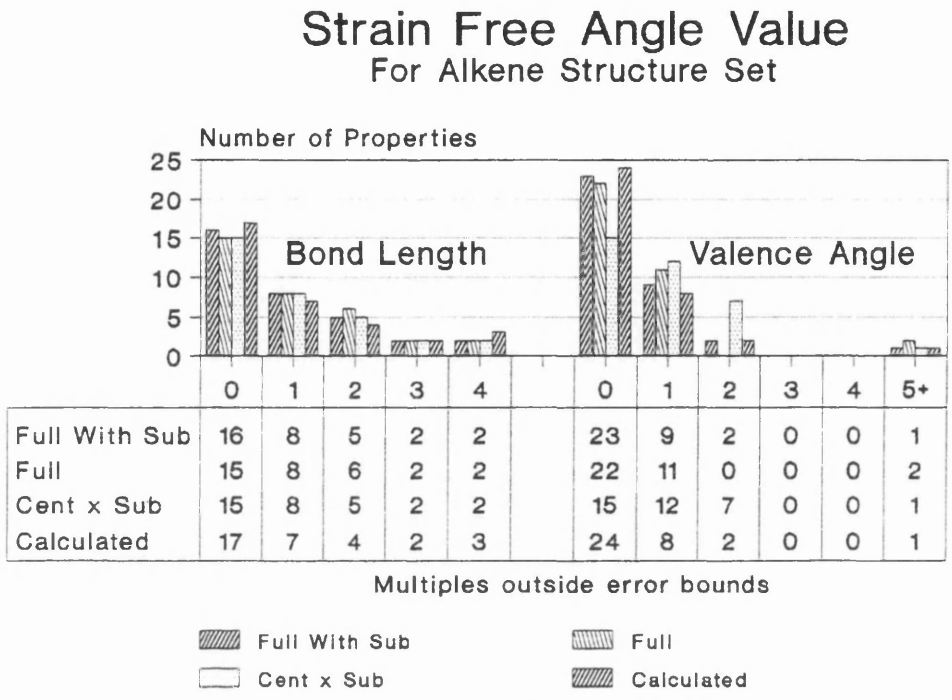


	Full with Sub	Calculated	Full	Cent x Sub	Cent only
Total	0.29	0.372	0.411	0.429	0.462
Heat of Formation	0.034	0.09	0.096	0.112	0.122
Conformer	0	0.003	0.001	6.800E-04	8.300E-04
Structure	0.257	0.28	0.313	0.316	0.339

This shows that the best results were obtained using the 'full with substitution' option, as would be expected. The option that comes second is not the 'full' case, which has the next largest number of parameters, but is instead the 'calculated' option. The main difference between all the options is the heat of formation residual factor for the 'full with substitution' option that is approximately a third of its value for the other options.

It is possible, as stated above, to process the COMPARE files for these options to investigate the results of the optimisations in greater detail. Unfortunately the COMPARE file for the 'central only' case has been lost since the optimisation took place. But, as this showed by far the worst total residual factor, and involves only two fewer parameters that the 'central times substitution factor' option this loss is not significant. Processing of the remaining COMPARE files gives rise to the residual factor distributions shown in figure 7.15. Figure 7.16 then shows the standard deviation for these four cases.

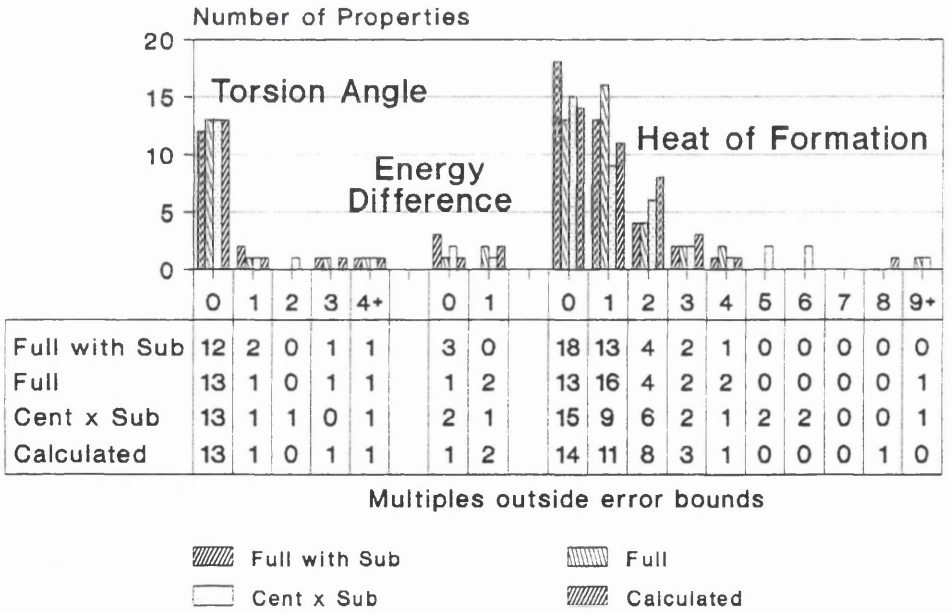
Figure 7.15 Residual Factors for Strain Free Angles, Alkene Study



	Bond Length					Valence Angle					
	0	1	2	3	4	0	1	2	3	4	5+
Full With Sub	16	8	5	2	2	23	9	2	0	0	1
Full	15	8	6	2	2	22	11	0	0	0	2
Cent x Sub	15	8	5	2	2	15	12	7	0	0	1
Calculated	17	7	4	2	3	24	8	2	0	0	1

Multiples outside error bounds

Full With Sub
Full
Cent x Sub
Calculated

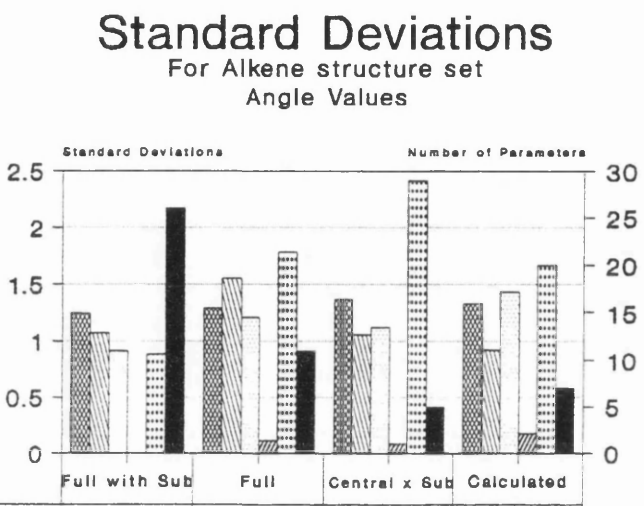


	Torsion Angle					Energy Difference		Heat of Formation								
	0	1	2	3	4+	0	1	0	1	2	3	4	5	6	7	8	9+
Full with Sub	12	2	0	1	1	3	0	18	13	4	2	1	0	0	0	0	0
Full	13	1	0	1	1	1	2	13	16	4	2	2	0	0	0	0	1
Cent x Sub	13	1	1	0	1	2	1	15	9	6	2	1	2	2	0	0	1
Calculated	13	1	0	1	1	1	2	14	11	8	3	1	0	0	0	1	0

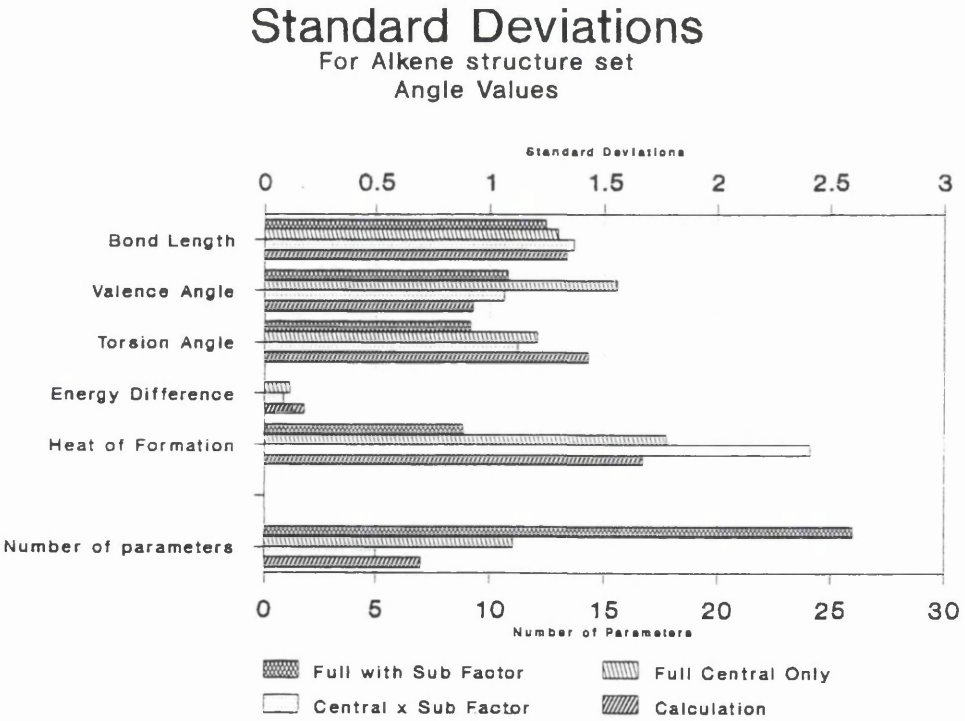
Multiples outside error bounds

Full with Sub
Full
Cent x Sub
Calculated

Figure 7.16 Standard Deviations for Strain Free Angles, Alkene Study



	Full with Sub	Full	Central x Sub	Calculated
Bond Lengths	1.242	1.296	1.367	1.332
Valence Angles	1.077	1.558	1.061	0.926
Torsion Angles	0.913	1.204	1.12	1.429
Energy Differences	0	0.114	0.085	0.178
Heat of Formation	0.882	1.778	2.414	1.669
Number of Parameters	26	11	5	7



These results confirm those from the comparison of the residual factor. They also bring to light a strange anomaly in the 'full' case that has two bad angles compared to one for the other options. The reason for this is unknown and the result is that the standard deviation is worse than would be expected by examining the distribution of the valence angles nearer to zero. This shows that the 'full' option is better than the 'central times substitution factor' option but even ignoring the dubious bad angle it is not as good as the 'calculated' option.

The problem with the 'full with substitution' method is that of a deficient structure set. As can be seen from Figure 7.15 there are 35 valence angles specified in the Alkene structure set. As stated in Figure 7.16 the 'full with substitution' used 26 separate parameters. Because it is difficult to get an even distribution of the angles' types within the structure this means that there are several parameters that are not directly related to the experimental data. This will result in parameters that have optimised only to non-structural data, such as heat of formation, and so become unreliable. For example, in the optimisation for which the data is shown, the C_{sp3}-C_{sp3}-C_{sp3} angle with only one hydrogen attached optimised to a value of 103.8°, whereas the same angle but with two hydrogens optimises to 109.2°.

If the object were to produce an alkene only force field then it is likely that the best situation would be to increase the number of experimental angles defined in the structure set and to use the 'full with substitution' option, as the number of parameters would still be manageable. This, though, is not the final objective and so the next step was to investigate these options for the expanded structure set to examine what effect the extra atom types have.

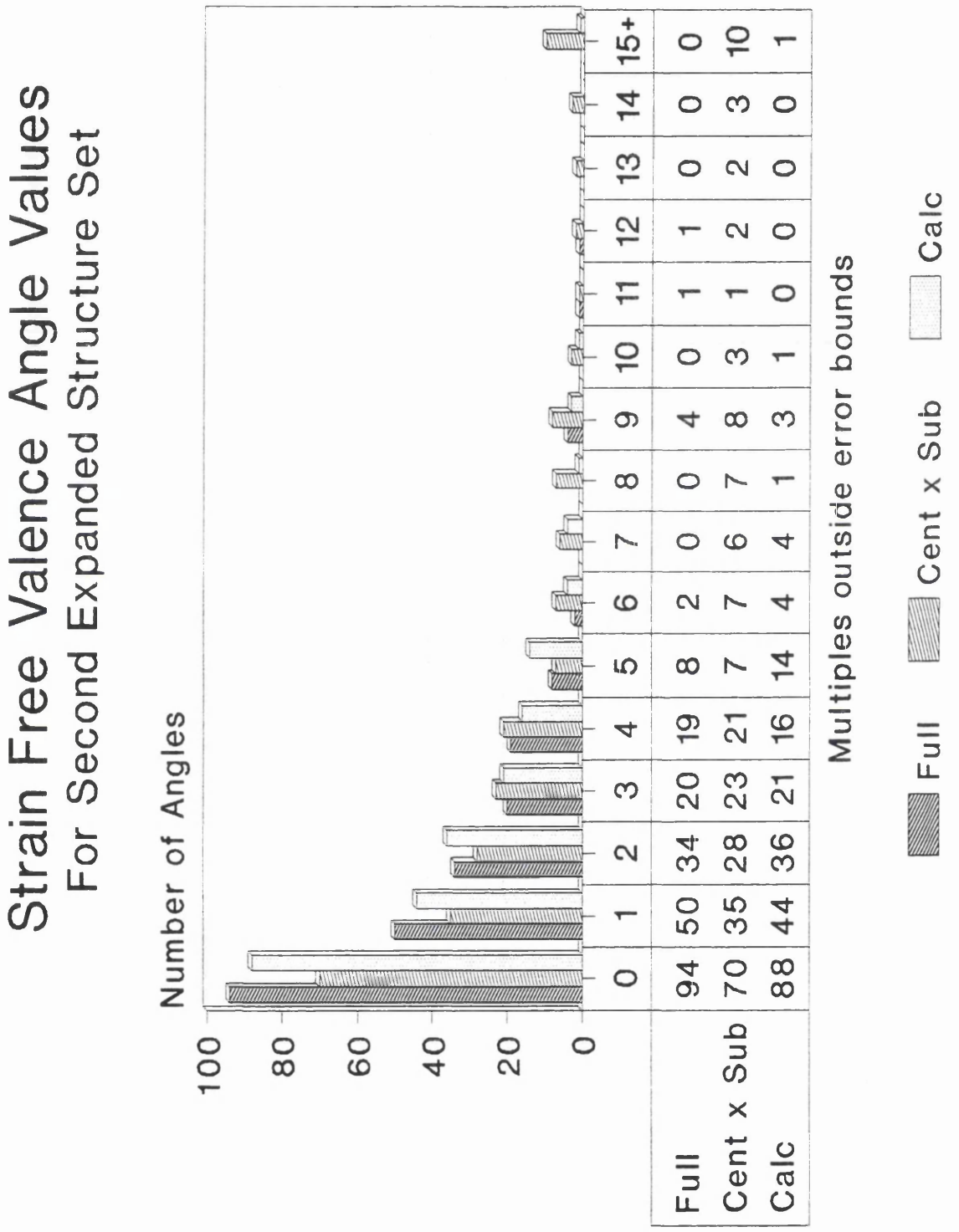
7.4.2.1.2 Expanded Structure Set

The same approach was made with the second expanded structure set. Figure 7.17 show the distribution of the residual factors for the Valence angles in these cases. Figures 7.18 and 7.19 show the distribution of the other properties. For the expanded structure set a 'full with substitution' option was not used as it was decided that the structures used provided insufficient depth over the whole range of possible angles to produce meaningful results.

Comparing the remaining options it can be seen that the 'full' case gives the best results, but that in all residual factors, except heat of formation, it is closely followed by the 'calculated' option.

The worst case can be seen to be the "central times substitution factor" situation. The reason for this is that, though the use of a substitution count scaling factors may be reasonable when applied to an alkene force field when all the non-Hydrogen's are Carbons, the same cannot be said for a force field with more atom

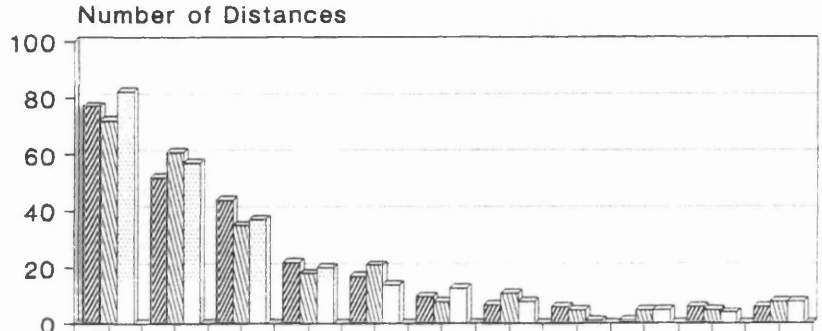
Figure 7.17 Valence Angle Residual Factors for Strain Free Angle Options



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
Full	94	50	34	20	19	8	2	0	0	4	0	1	1	0	0	0
Cent x Sub	70	35	28	23	21	7	7	6	7	8	3	1	2	2	3	10
Calc	88	44	36	21	16	14	4	4	1	3	1	0	0	0	0	1

Figure 7.18 Residual Factors for Other Properties (I)

Distances for Strain Free Valence Angles
For Second Expanded Structure Set

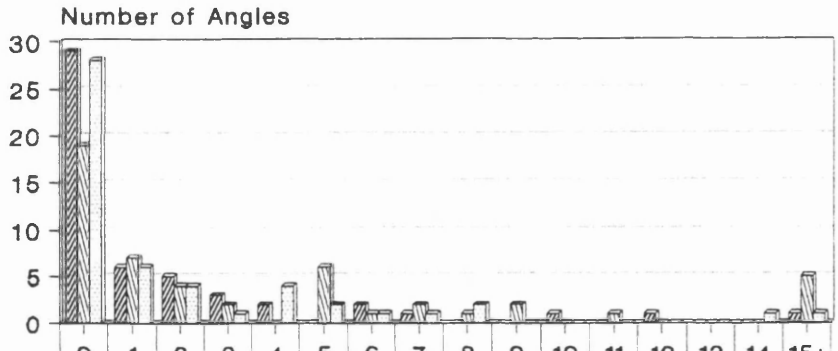


	0	1	2	3	4	5	6	7	8	9	10+
Full	77	52	44	22	17	10	7	6	1	6	6
Central x Sub	72	61	35	18	21	8	11	5	5	5	8
Calculated	82	57	37	20	14	13	8	1	5	4	8

Multiples outside error bounds

Full Central x Sub Calculated

Torsion Angles for Strain Free Angles
For Second Expanded Structure Set

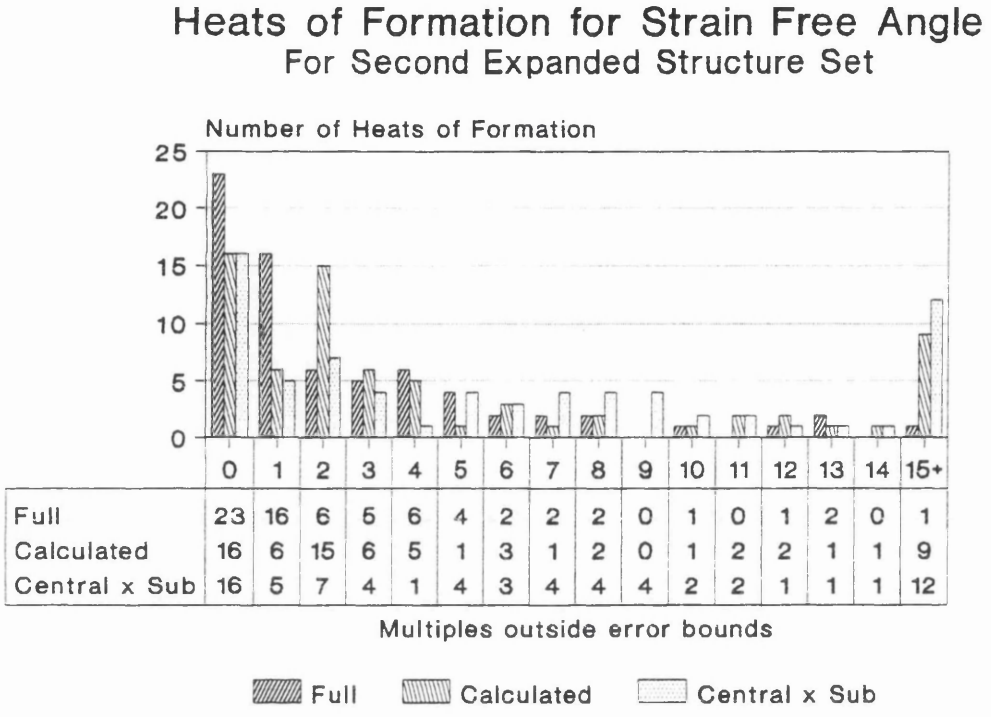


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
Full	29	6	5	3	2	0	2	1	0	0	1	0	1	0	0	1
Cent x Sub	19	7	4	2	0	6	1	2	1	2	0	1	0	0	0	5
Calc	28	6	4	1	4	2	1	1	2	0	0	0	0	0	1	1

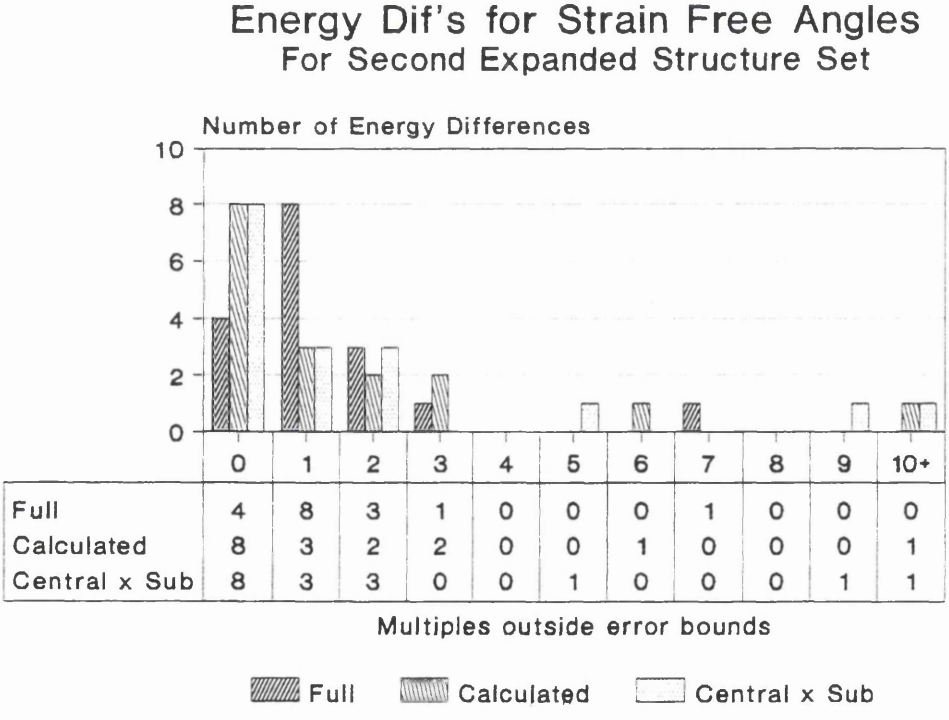
Multiples outside error bounds

Full Cent x Sub Calc

Figure 7.19 Residual Factors for Other Properties (II)

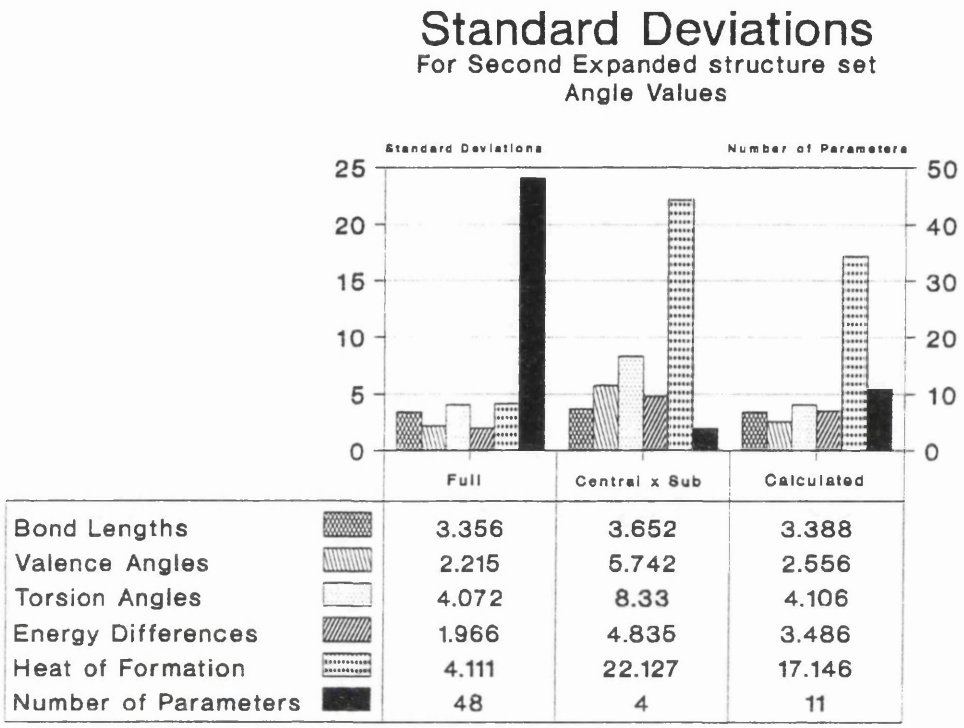


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
Full	23	16	6	5	6	4	2	2	2	0	1	0	1	2	0	1
Calculated	16	6	15	6	5	1	3	1	2	0	1	2	2	1	1	9
Central x Sub	16	5	7	4	1	4	3	4	4	4	2	2	1	1	1	12

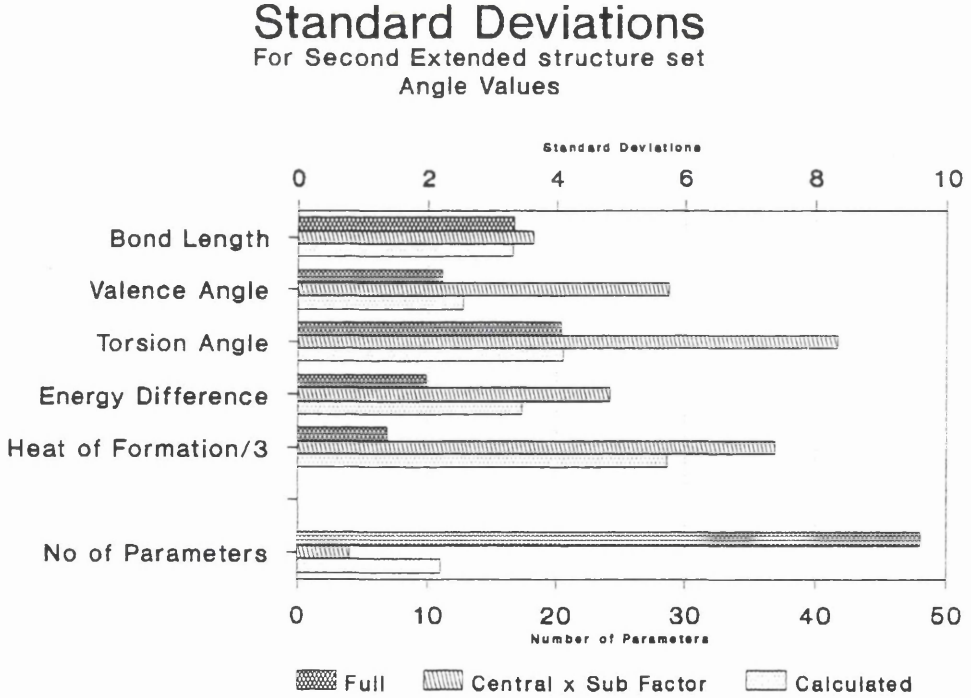


	0	1	2	3	4	5	6	7	8	9	10+
Full	4	8	3	1	0	0	0	1	0	0	0
Calculated	8	3	2	2	0	0	1	0	0	0	1
Central x Sub	8	3	3	0	0	1	0	0	0	1	1

Figure 7.20 Standard Deviations for Strain Free Angle Options



	Full	Central x Sub	Calculated
Bond Lengths	3.356	3.652	3.388
Valence Angles	2.215	5.742	2.556
Torsion Angles	4.072	8.33	4.106
Energy Differences	1.966	4.835	3.486
Heat of Formation	4.111	22.127	17.146
Number of Parameters	48	4	11



types. In those cases it is unreasonable, for example, to assume that the strain free angle for H-C_{sp3}-H in CBr_2H_2 will be the same as that in CF_2H_2.

Figure 7.20 shows the standard deviations for the investigated cases. The major difference is that the heat of formation residual factor is much lower in the 'full' option as compared to the other two. This follows the general trend for those options with many parameters and again it is likely to be caused by the optimisation of parameters not sufficiently covered in the structural data that deform to optimise the energy residual factors more than would be expected for the 'global' structure set.

Again the problem of choosing the optimum case is one of balance. To help decide the best case it is pertinent to estimate the number of parameters that would be required if each option were to be used for a force field with a greater number of atom types. The numbers of parameters required to fully define the force field for all possible cases are given in table 7.3.

Table 7.3 Comparison of Parameter Numbers for Strain Free Angle Options

Full with Substitution Factor	≈6000
Full	≈2000
Calculated	47
Central x Substitution Factor	18

Thus to produce a force field with the 'full' option not only would 2000 parameters have to be optimised, but more significantly, the structure set would need an unwieldy number of structures to ensure it was not deficient for any parameter.

If the 'full' option is not to be used then the choice is between the 'central times substitution' option and the 'calculated' option. Compared to the 'full' case there is little difference between the required number of parameters in these cases, with both requiring a relatively small number of parameters. In the tables of standard deviations it can be seen that the 'calculated' option is far better than the 'central times substitution' option for all the residual factors except the distances, where there is almost no significant difference over all the optimisations. Thus on this data the optimum situation for the strain free angle is taken as the 'calculated' option.

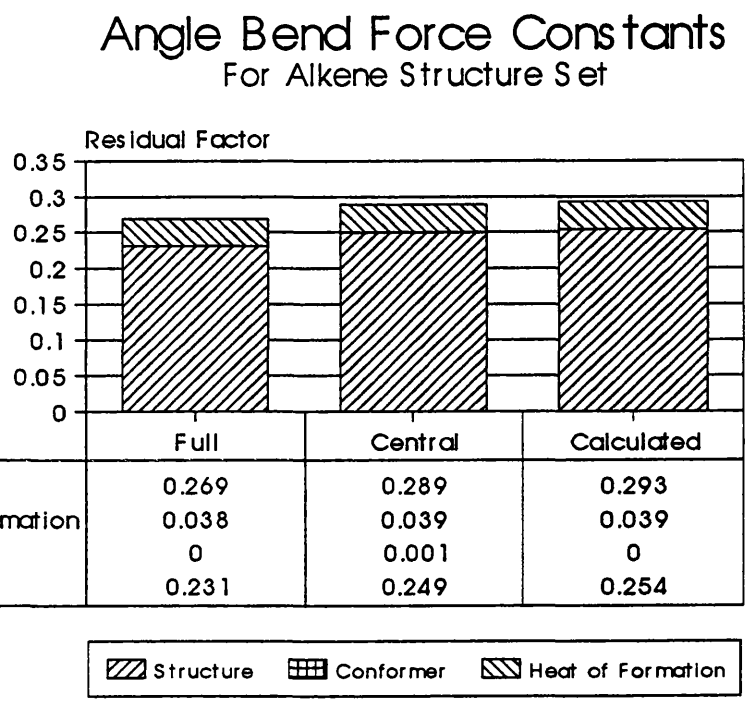
7.4.2.2 Angle Bending Force Constant

7.4.2.2.1 Alkene Structure Set

For this case three situations were investigated for the Alkene structure set. These were the full case that involved a separate force constant for each possible angle, central only where the force constant was decided depending only on the

central atom type and the calculated case with its dependence on bond lengths and electronegativities. Figure 7.21 shows the total residual factors in these cases.

Figure 7.21 Total Residual Factors for Angle Bend Options, Alkene Study



	Full	Central	Calculated
Total	0.269	0.289	0.293
Heat of Formation	0.038	0.039	0.039
Conformer	0	0.001	0
Structure	0.231	0.249	0.254

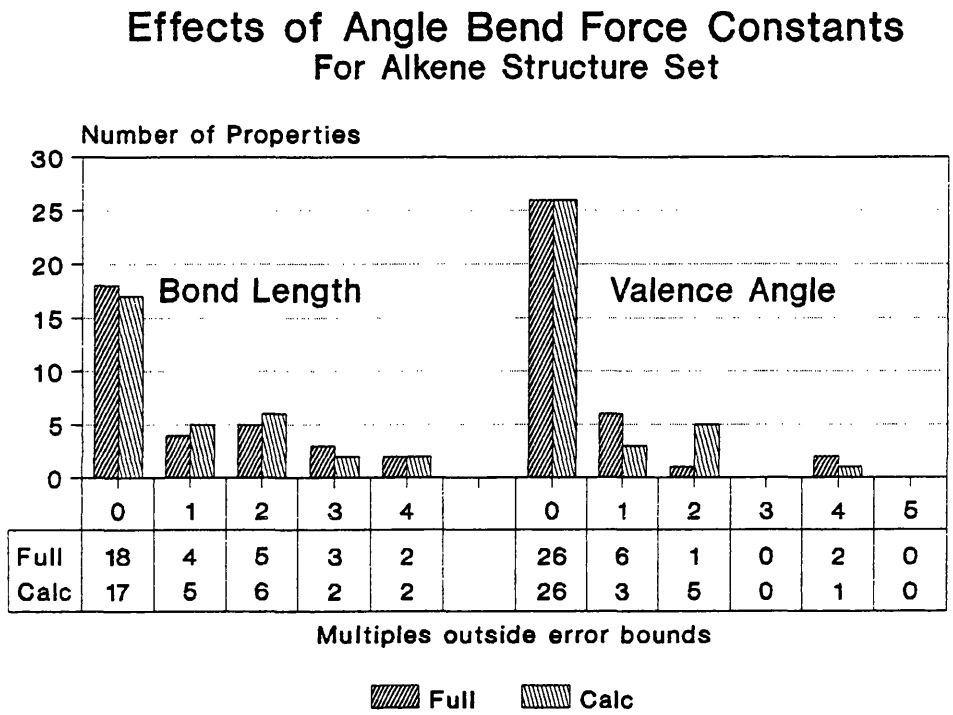
It can be seen from this that the only residual factor significantly separating the options is the structure. Unfortunately between the collection of the data and its analysis the COMPARE file for the central only case has been lost. Figure 7.22 shows the residual factor distribution for the remaining cases. Again it can be seen how close the two approaches are.

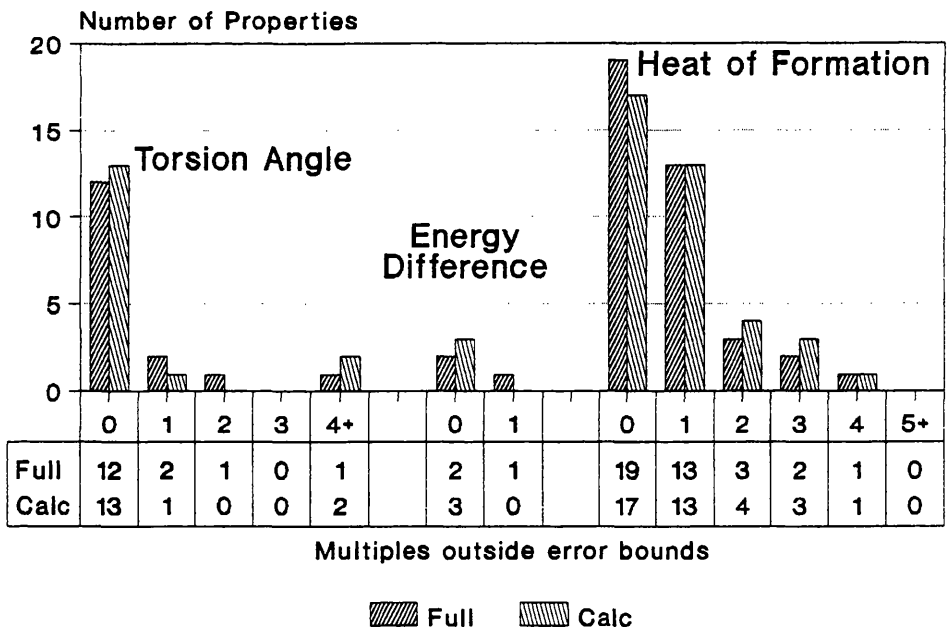
With such little difference between the angle bend methods for the alkene structure set it is necessary to investigate the various approachs with the expanded structure set before a decision can be made.

7.4.2.2.2 Expanded Structure Set

The angle bending case is unusual in that more methods were investigated with the expanded set than with the alkene structure set. The reason for this is that it was thought interesting to see if the use of a hydrogen substitution factor similar to that used in one of the angle value approaches would have a significant effect on the resultant residual factors. The distribution of the valence angles for the four cases under investigation is given in Figure 7.23. The distributions of the other properties are given in the graphs in figures 7.24 and 7.25. The summary of the standard deviations for these cases is given in figure 7.26.

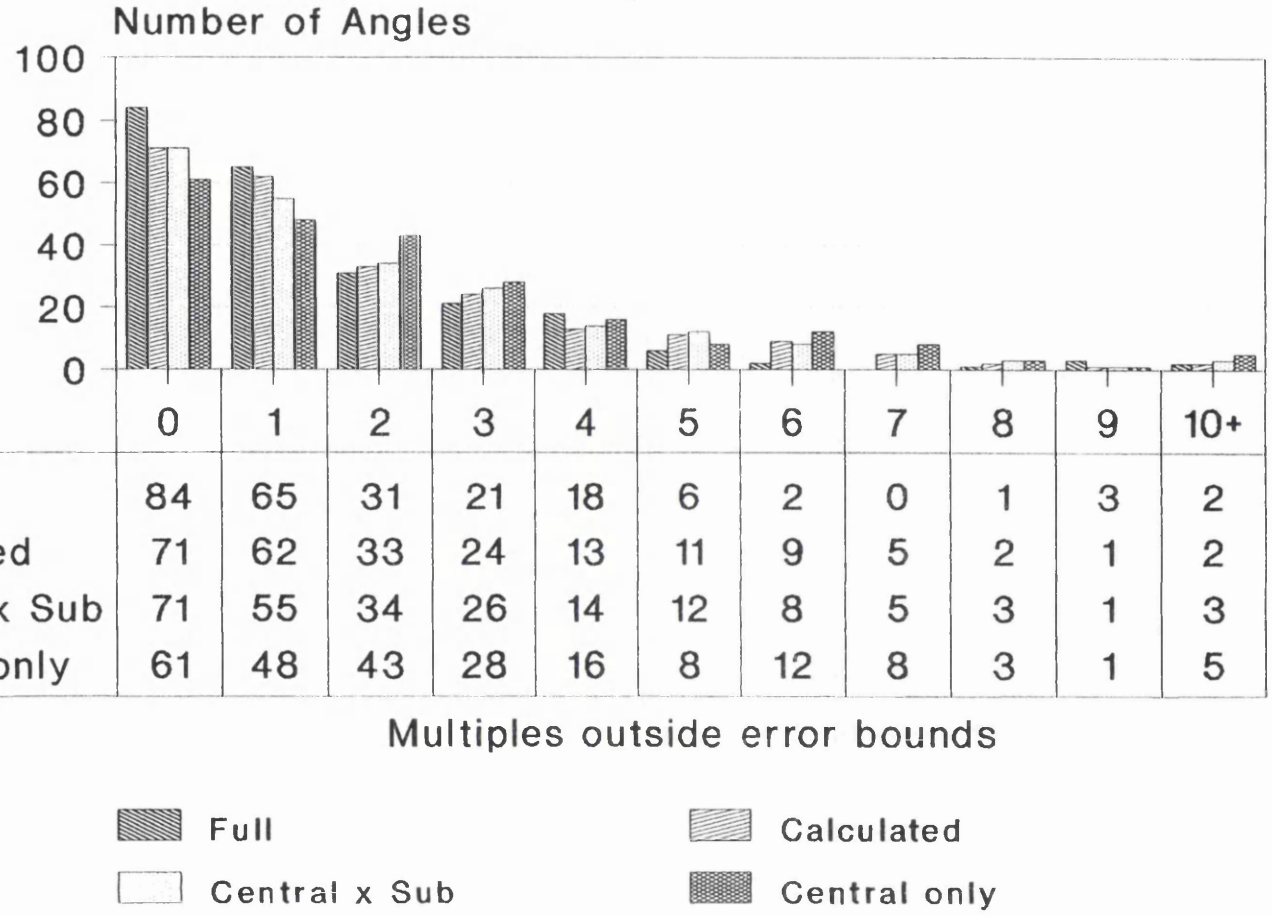
Figure 7.22 Residual Factor Chart for Angle Bend Options, Alkene Study





Angle Bend Force Constants

For Second Expanded Structure Set

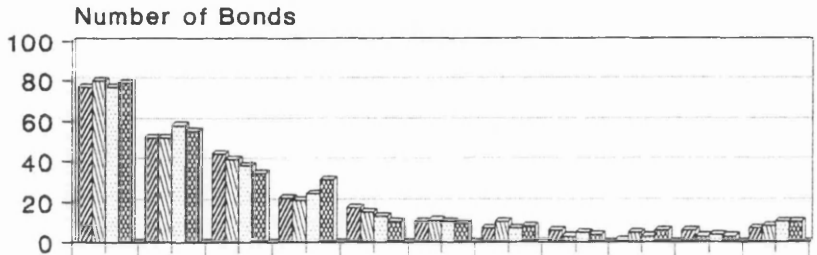


	0	1	2	3	4	5	6	7	8	9	10+
Full	84	65	31	21	18	6	2	0	1	3	2
Calculated	71	62	33	24	13	11	9	5	2	1	2
Central x Sub	71	55	34	26	14	12	8	5	3	1	3
Central only	61	48	43	28	16	8	12	8	3	1	5

Figure 7.23 Valence Angle Residual Factors For Angle Bend Options

Figure 7.24 Residual Factors for Other Options (I)

Lengths for Angle Bend Force Constants
For Second Expanded Structure Set

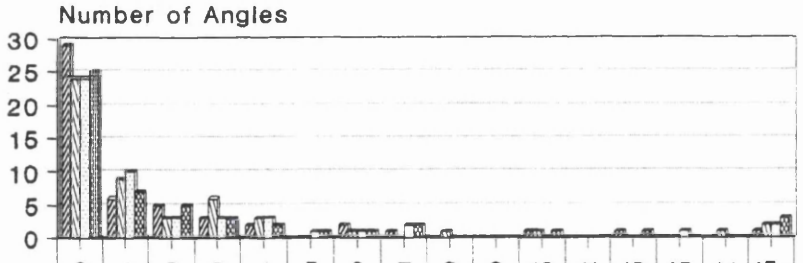


	0	1	2	3	4	5	6	7	8	9	10+
Full	77	52	44	22	17	10	7	6	1	6	7
Calculated	80	52	41	21	15	11	10	3	5	3	8
Central x Sub	77	58	38	24	13	10	7	5	3	4	10
Central only	79	55	34	31	10	9	8	4	6	3	10

Multiples outside error bounds

Full Calculated
Central x Sub Central only

Torsion Angles for Angle Bend Constants
For Second Expanded Structure Set



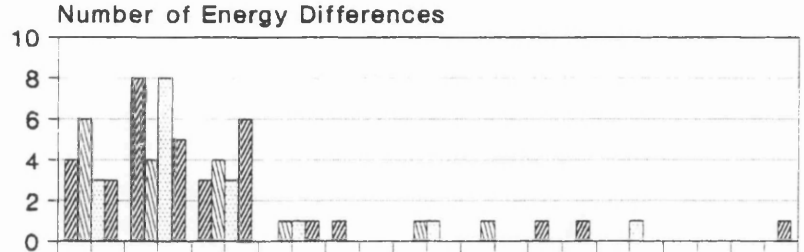
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
Full	29	6	5	3	2	0	2	1	0	0	1	0	1	0	0	1
Calculated	24	9	3	6	3	0	1	0	1	0	1	0	0	0	1	2
Central x Sub	24	10	3	3	3	1	1	2	0	0	0	0	0	1	0	2
Central only	25	7	5	3	2	1	1	2	0	0	1	0	1	0	0	3

Multiples outside error bounds

Full Calculated
Central x Sub Central only

Figure 7.25 Residual Factors for Other Options (II)

Energy Dif's for Angle Bend Force Consts
For Second Expanded Structure Set

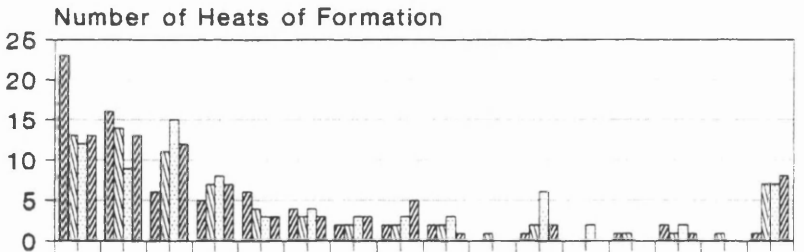


	0	1	2	3	4	5	6	7	8	9	10+
Full	4	8	3	0	1	0	0	1	0	0	0
Calculated	6	4	4	1	0	1	1	0	0	0	0
Central x Sub	3	8	3	1	0	1	0	0	1	0	0
Central only	3	5	6	1	0	0	0	1	0	0	1

Multiples outside error bounds

Full Calculated
Central x Sub Central only

Heats of Formation for Angle Bend
For Second Expanded Structure Set

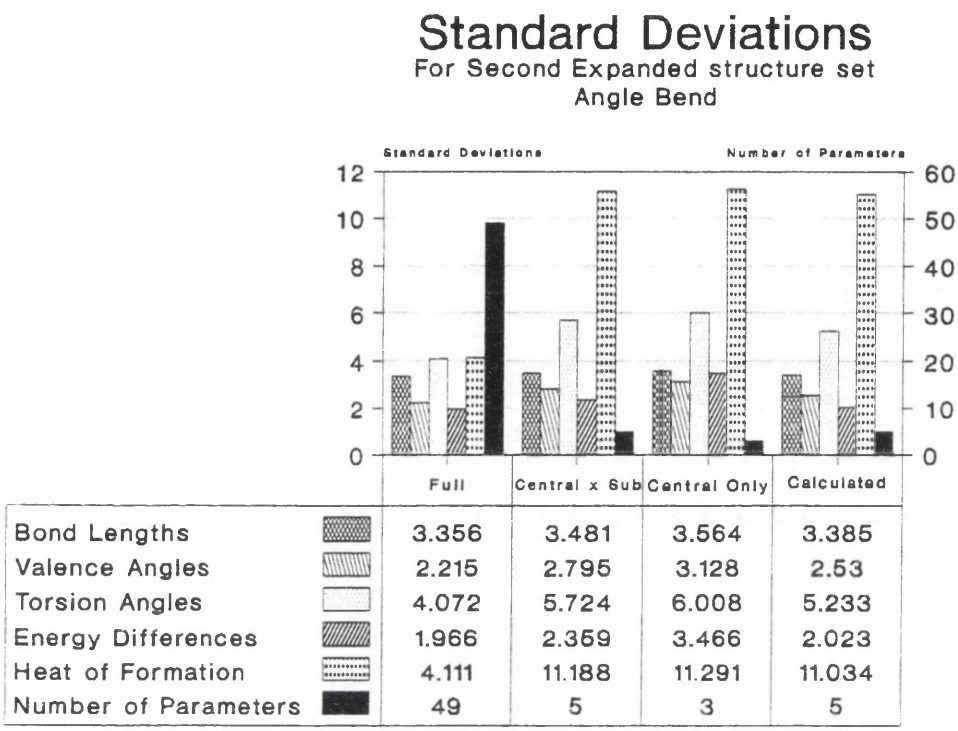


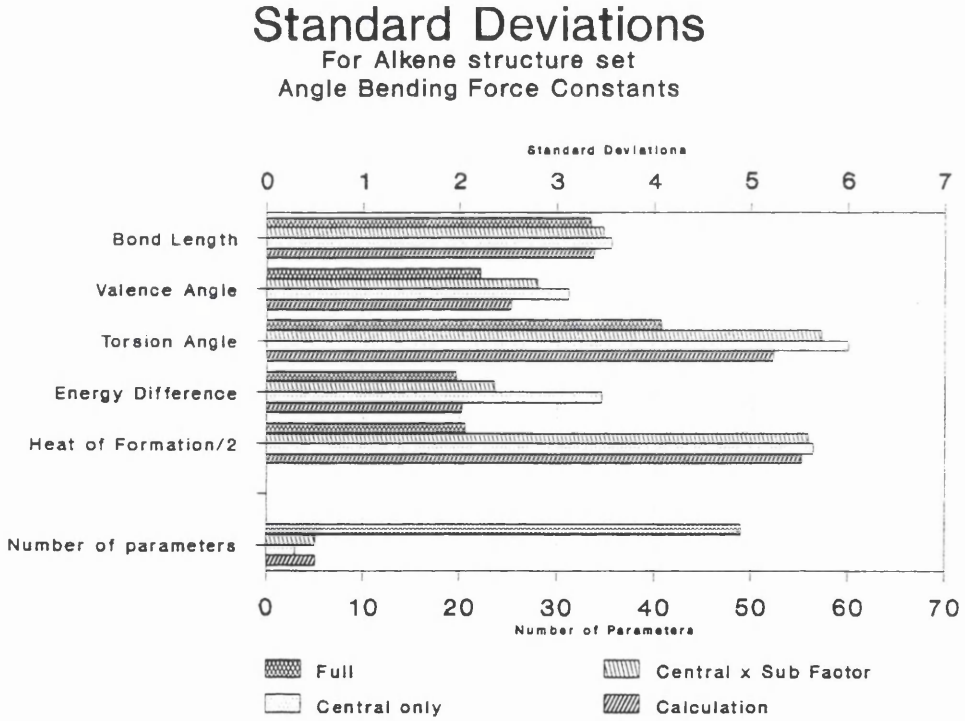
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
Full	23	16	6	5	6	4	2	2	2	0	1	0	1	2	0	1
Calculated	13	14	11	7	4	3	2	2	2	1	2	0	1	1	1	7
Central x Sub	12	9	15	8	3	4	3	3	3	0	6	2	0	2	0	7
Central only	13	13	12	7	3	3	3	5	1	0	2	0	0	1	0	8

Multiples outside error bounds

Full Calculated
Central x Sub Central only

Figure 7.26 Standard Deviations of Residual Factors for Angle Bend Options





Examination of the charts shows a very similar trend to those earlier. The lowest standard deviation in all the cases is for the 'full' option. Again the residual factors for the distances are almost identical and the standard deviation for the heat of formation residual factors is significantly lower for the 'full' case. There is no reason to believe that the reasons for these observations are any different to those given above for the strain free angle options.

As before the next step is to examine the number of parameters that would be required for a fuller force field for each of the options under investigation. These figures are given in table 7.4.

Table 7.4 Comparison of the Number of Parameters for Angle Bend Options

Full	≈2000
Central x Substitution Factor	≈17
Central Only	≈15
Calculated	18

Again the number of parameters required for the 'full' option is significantly larger and would be unworkable if the parameters were to be optimised using OPTPARAM. Of course it is likely that many of the possible angles will not turn up very often, if at all, and so fewer parameters need to be optimised to be part of a usable force field. Having said this, the advantage of all the other approaches is that once a small number of parameters have been optimised all possible configurations are then covered. Thus, again, the 'full' scheme cannot be chosen at the optimum option.

Unlike the strain free angle option the choice of the next best option is not obvious as all the standard deviations are too close to make it impossible to see the best option. This is especially true as all remaining options require only a few parameters compared to the fuller force field. Examining the standard deviations the information in table 7.5 can be obtained.

Table 7.5 Selected Standard Deviation Data for Angle Bend Options

Option	Average S.D.	Number of Lowest S.D's
Central x Sub Factor	5.109	0
Central Only	5.491	0
Calculated	4.841	5

Thus if the 'full' case is not chosen then the next best method is that using the

calculated angle bend force constants and this is the optimum method chosen for use in the final force field.

7.4.3 Non-Bonded Interactions

Non-bonded interactions are different from the other interactions in two ways. The first of these is that, as stated before, the choice is not so much between different ways of determining the parameters to be used but more between different forms of the interaction equations. The second reason is that there is no one structural property of the molecules involved in the study that will be the major one affected by the different methods. For example, it would be expected that bond lengths would be the main property affected by the bond stretch force constant parameters and that the valence angles would be the most critical properties for the strain free angles and the bending constants. There is no such obvious relationship for the non-bonded interactions and because of this it is important to consider all the different property types before deciding on the optimum method for each case.

7.4.3.1 Alkene Structure Set

The options investigated using the alkene structure set were:

A6-B9 A6-B12 A6-Exponential

Cosmic

Figure 7.27 shows the distribution of the residual factors for these four types of non-bonded interactions optimised with the Alkene structure set. It can be seen that the first three interactions' types are fairly equal with the COSMIC[1] non-bonded interaction type being only slightly worse in most cases. This is again consistent with the number of parameters being optimised in each case as the first 3 involve the optimisation of 12 parameters whereas the COSMIC case uses only 6 parameters.

Again it is impossible to decide the optimum type of interaction using the Alkene structure set only, so this must be decided after an investigation using the expanded structure set.

7.4.3.2 Expanded Structure Set

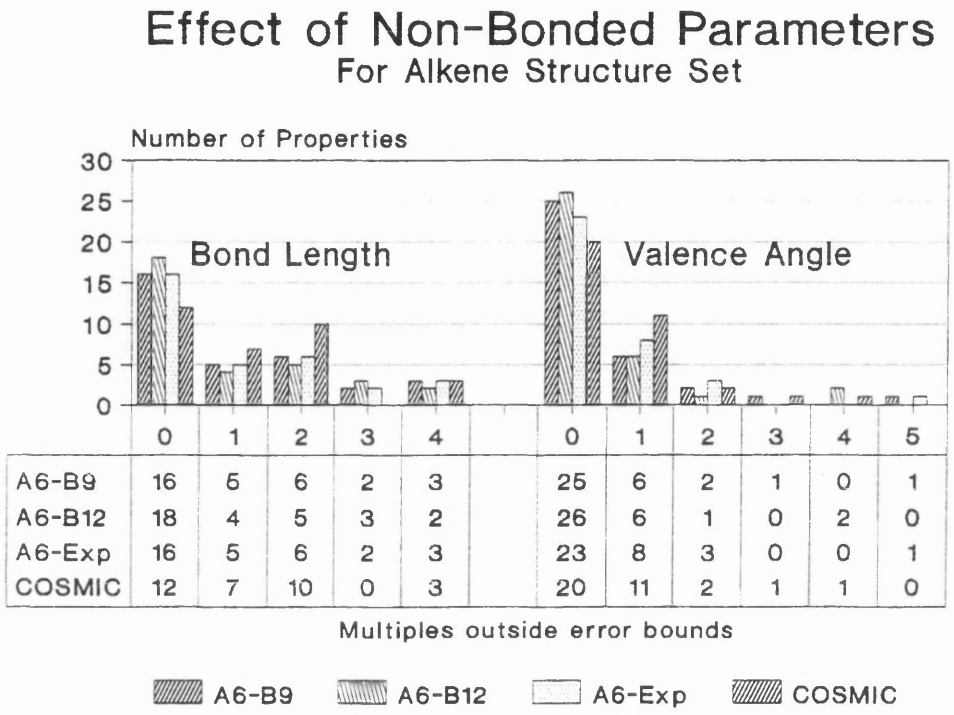
The largest range of non-bonded options was investigated using the first expanded structure set. These were:

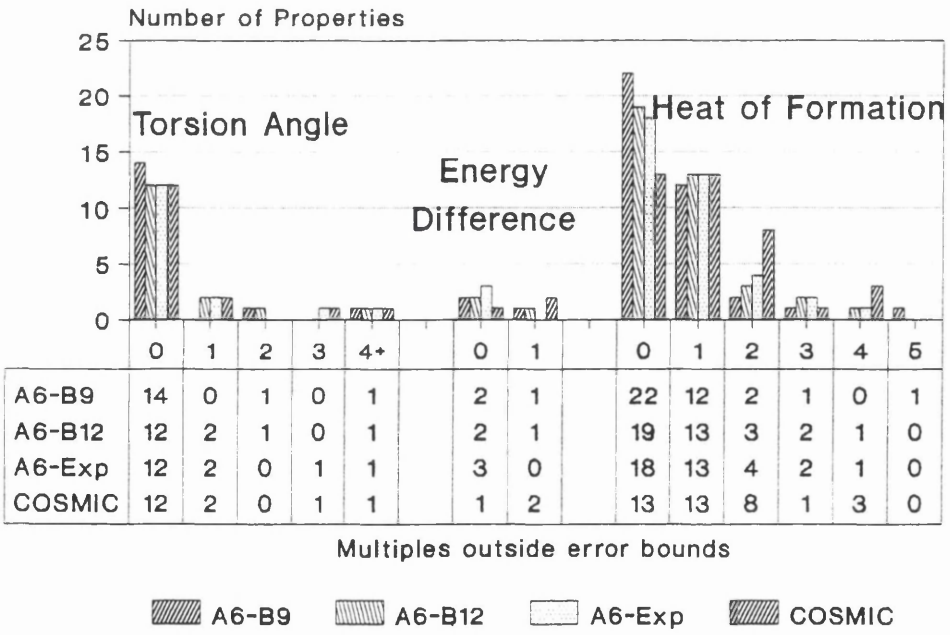
A6-B9 A6-B12 A6-Exponential

COSMIC A6-B9 Geometric Mean

Figures 7.28 to 7.32 show the distribution of the residual factors for the various properties and figure 7.33 shows the standard deviations for these cases.

Figure 2.27 Residual Factors for Non-Bonded Options, Alkene Study





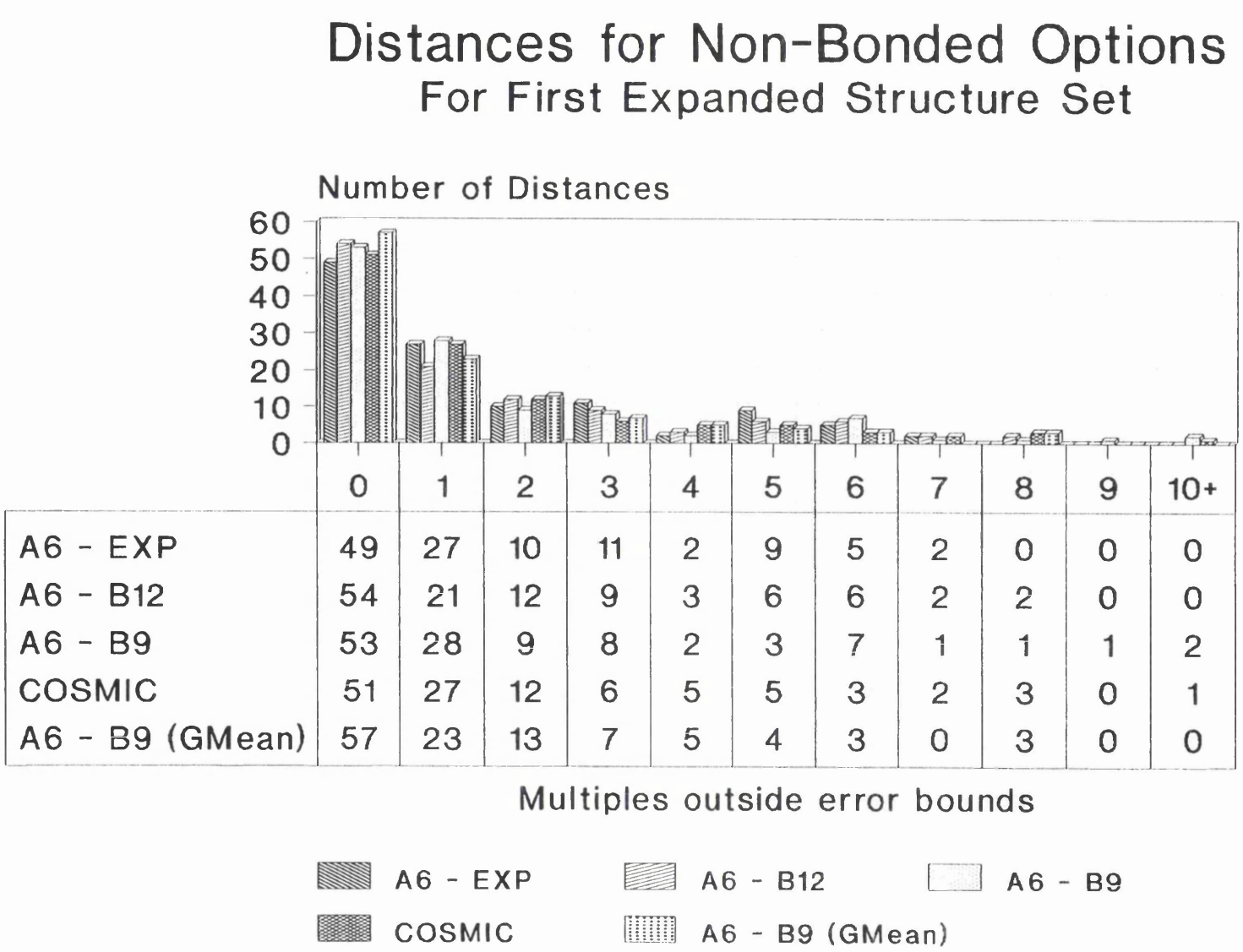


Figure 7.28 Distance Residual Factors for Non-Bonded's, 1st Extended

Figure 7.29 Valence Angle Residual Factors for Non-Bonded's, 1st Extended

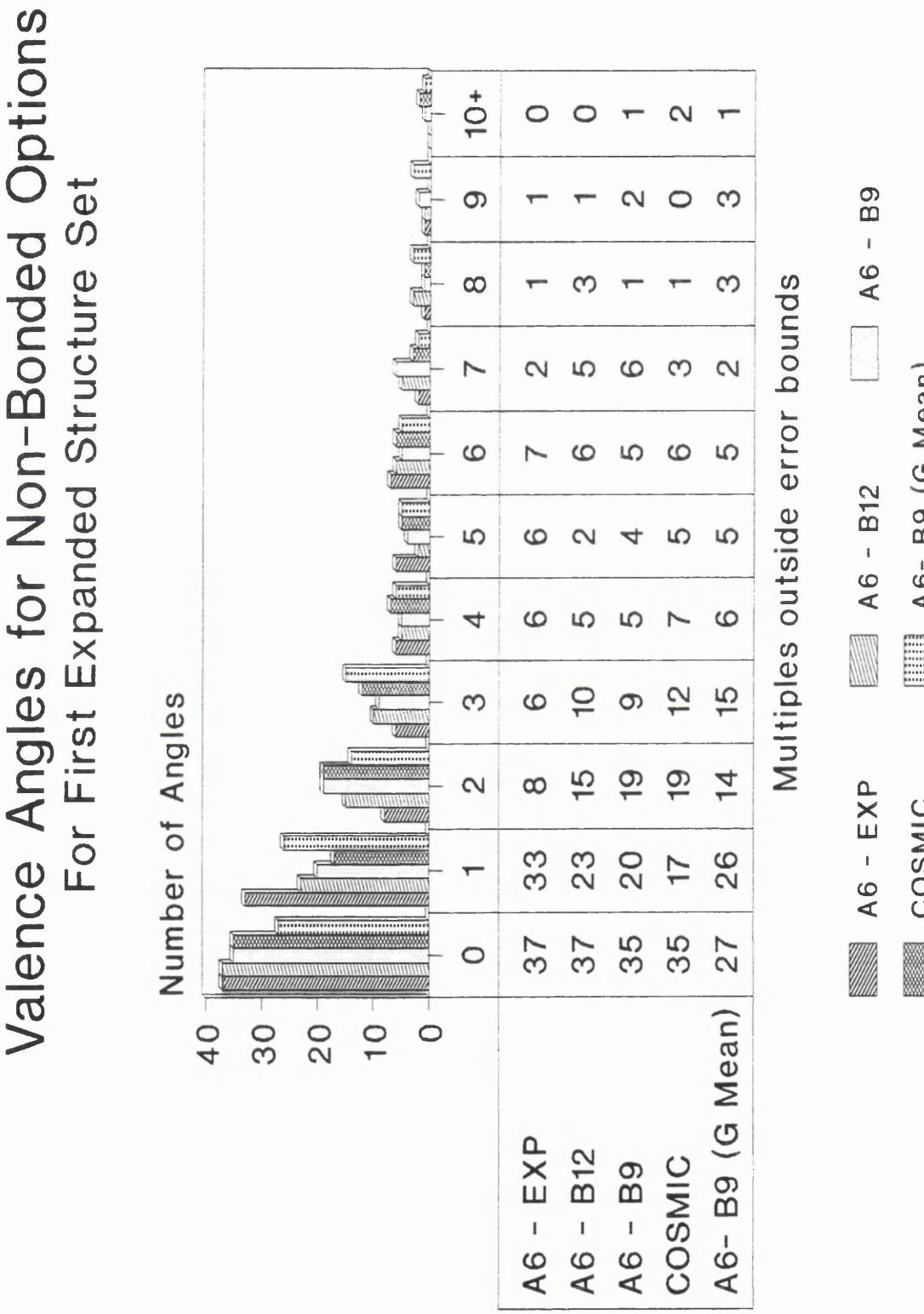


Figure 7.30 Torsion Angle Residual Factors for Non-Bonded's, 1st Extended

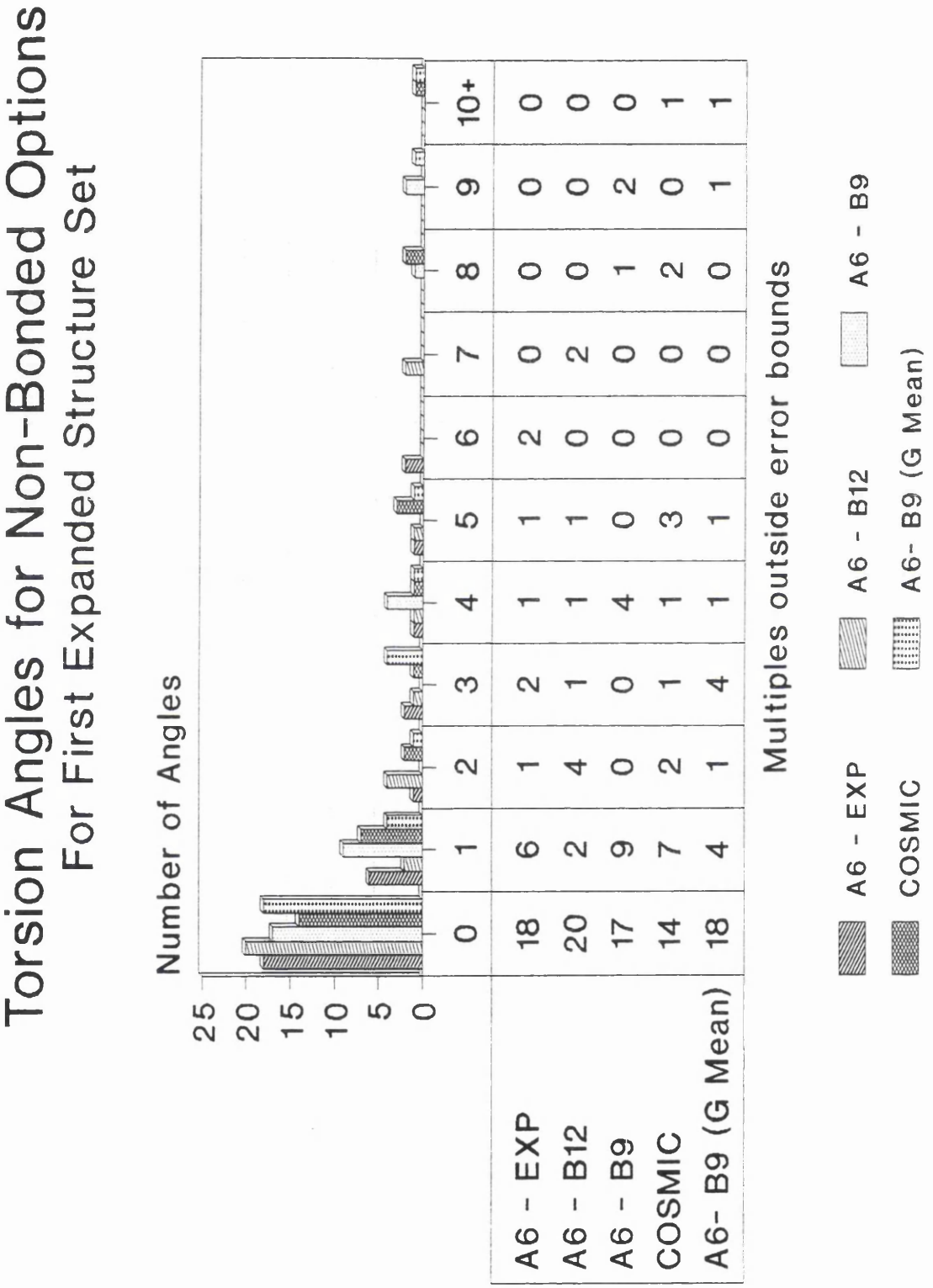


Figure 7.31 Isomer Energy Residual Factors for Non-Bonded's, 1st Extended

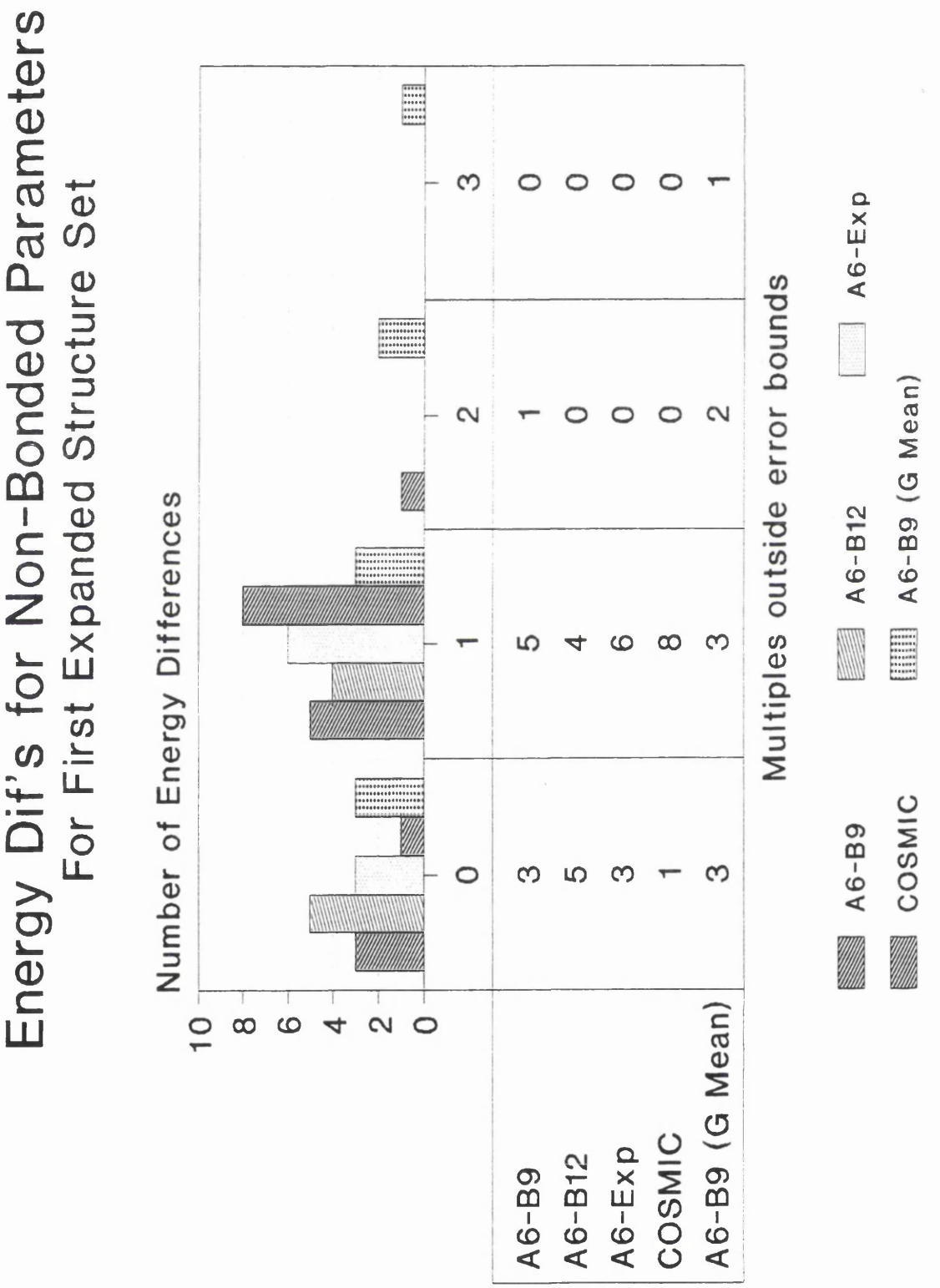
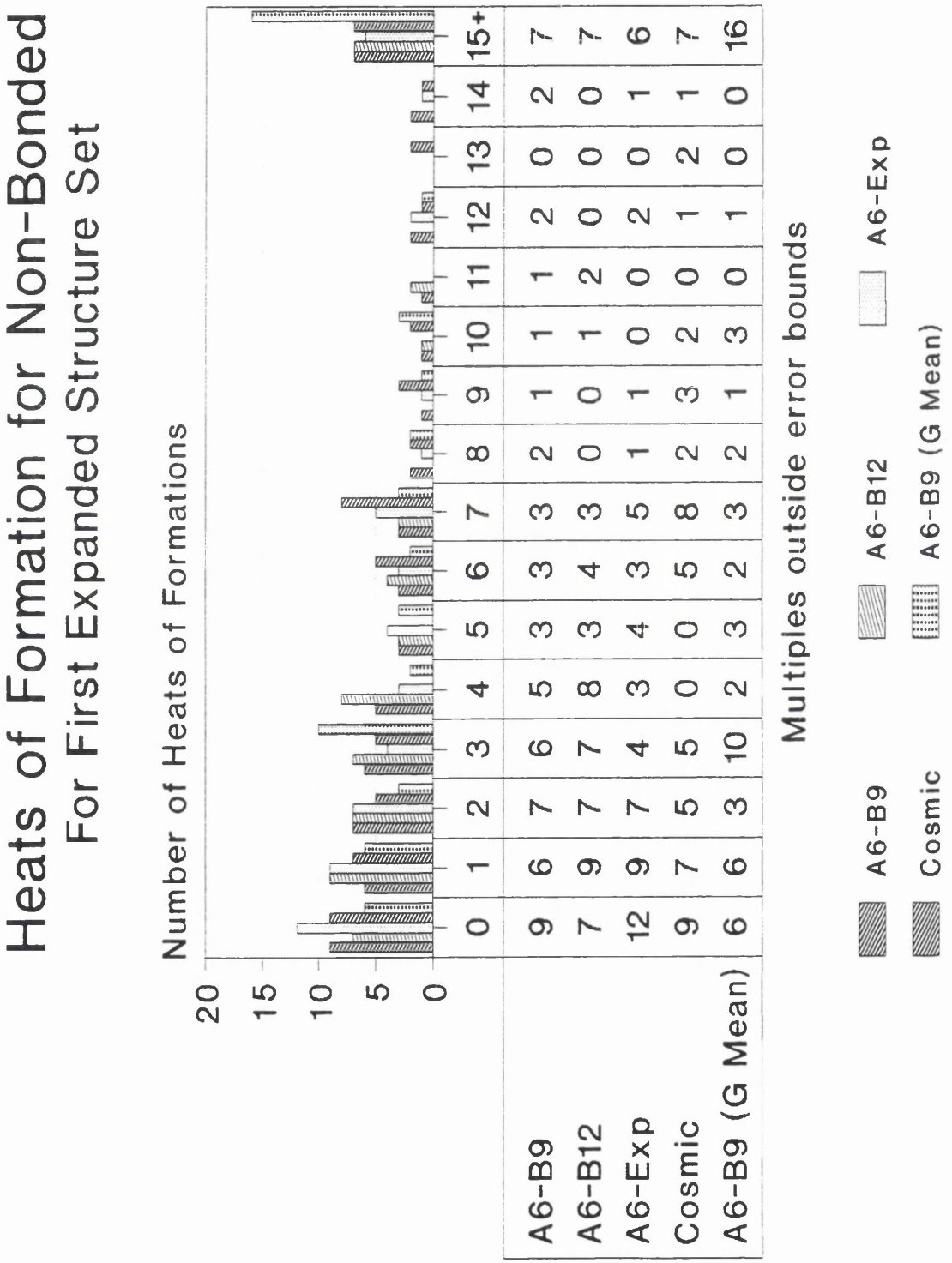
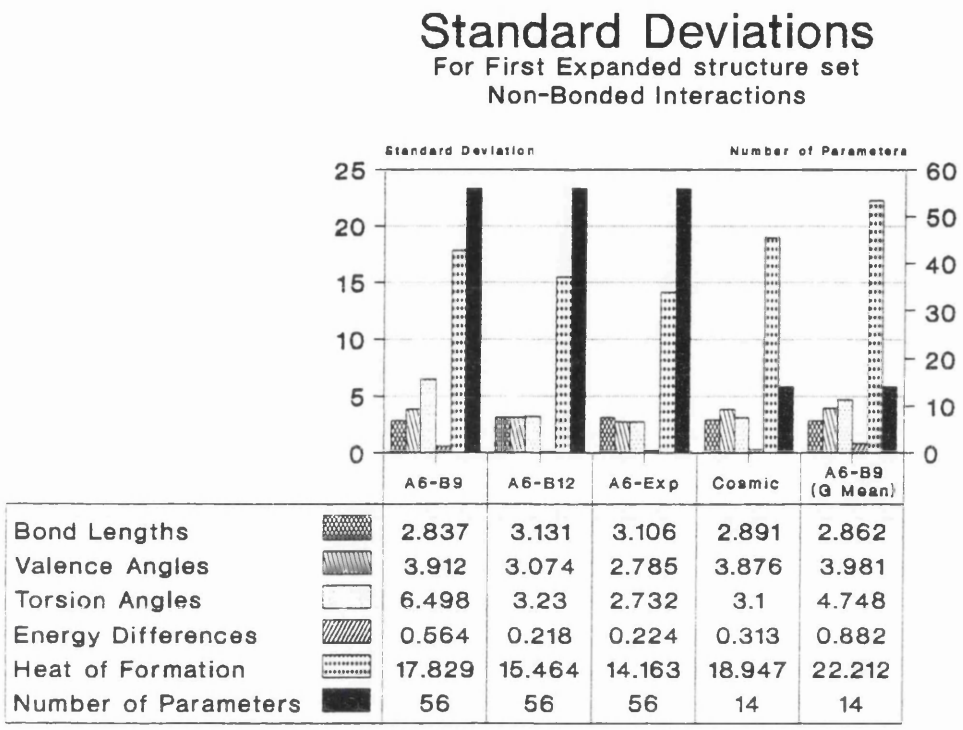


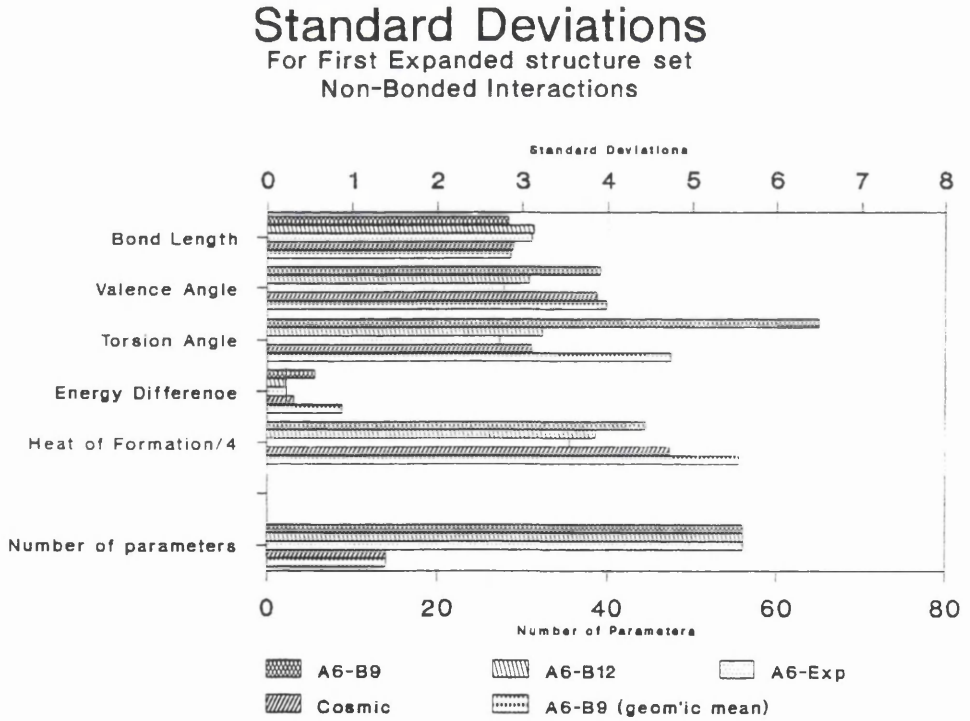
Figure 7.32 Heat of Formation Residual Factors for Non-Bonded's, 1st Extended



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
A6-B9	9	6	7	6	5	3	3	3	2	1	1	1	2	0	2	7
A6-B12	7	9	7	7	8	3	4	3	0	0	1	2	0	0	0	7
A6-Exp	12	9	7	4	3	4	3	5	1	1	0	0	2	0	1	6
Cosmic	9	7	5	5	0	0	5	8	2	3	2	0	1	2	1	7
A6-B9 (G Mean)	6	6	3	10	2	3	2	3	2	1	3	0	1	0	0	16

Figure 7.33 Standard Deviations for Non-Bonded Options, 1st Extended





As would be expected from the previous optimisations, examination of the distances cases shows that there is little difference between the investigations. This is also born out by the similar values for the standard deviations given in figure 7.32.

For the other residual factors it can be seen that the 'A6-Exp' option has the best results, having the lowest standard deviations in all cases except for the energy differences where it is very close to the lowest value produced by the 'A6-B12' option. After 'A6-Exp' the other options are fairly mixed: 'A6-B12' also has the next lowest valence angle and heat of formation residual factors but 'COSMIC' has the next lowest torsion angle value.

Overall it can be seen, from the information given, that of the full options the 'A6-Exp' is the best and that the 'COSMIC' option is significantly better than the other reduced parameter option, 'A6-B9 (geometric mean)'.

Thus, when the second expanded structure set was investigated, it was decided to try a further comparison of these two methods.

The following Figures 7.34 to 7.37 show the distribution of the residual factors for this comparison. Figure 7.38 shows the comparison of the standard deviations for these residual factors.

Examining the distributions and the standard deviations for each section.

Bond Length

As usual there is no significant difference in the bond length distribution between the two cases.

Valence Angle

Because the force field does not involve 1,3 non-bonded interactions the valence angles are not as affected as much as might be expected by the non-bonding interactions. In this case the distribution and the standard deviations are again similar between the two cases with the 'A6-Exp' case being slightly better.

Torsion Angle

Torsion angles are the structural properties most closely linked to the non-bonded interactions, as in many cases the final torsion angle location is reached from the balance of torsional twist and non-bonded interaction. For this residual factor it is found that the 'COSMIC' case is slightly better that the 'A6-Exp' case, though again the results are so close as to make it impossible to say which is definitively best.

Energy Differences

Again the graphs look fairly close with the exception that the 'COSMIC' option has two cases that are 4 or more times the error out from the experimental error. These two cases result in the standard deviation for the 'COSMIC' option being almost twice that for the 'A6-Exp' option.

Figure 7.34 Distance Residual Factors for Non-Bonded's, 2nd Extended

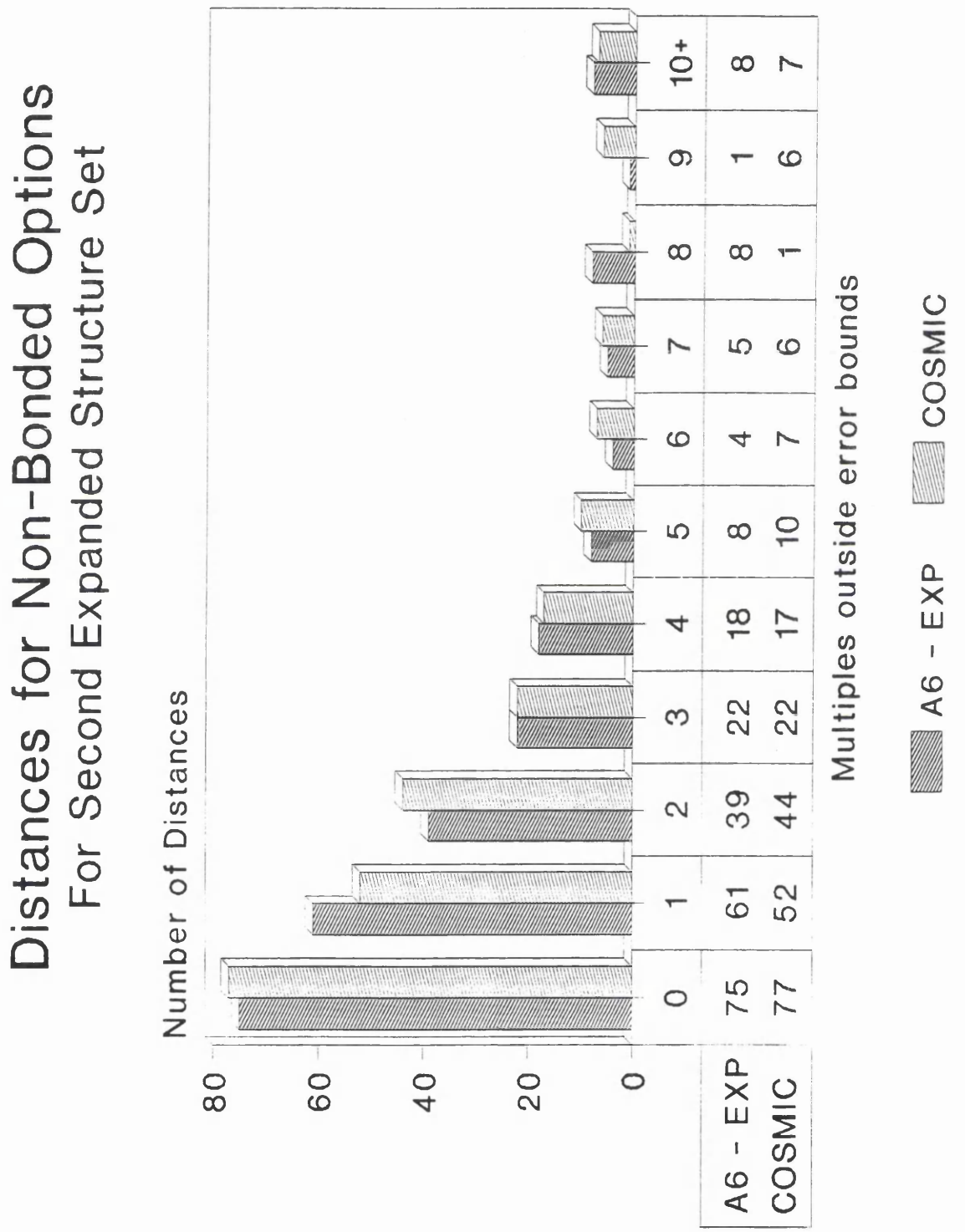


Figure 7.35 Valence Angle Residual Factors for Non-Bonded's, 2nd Extended

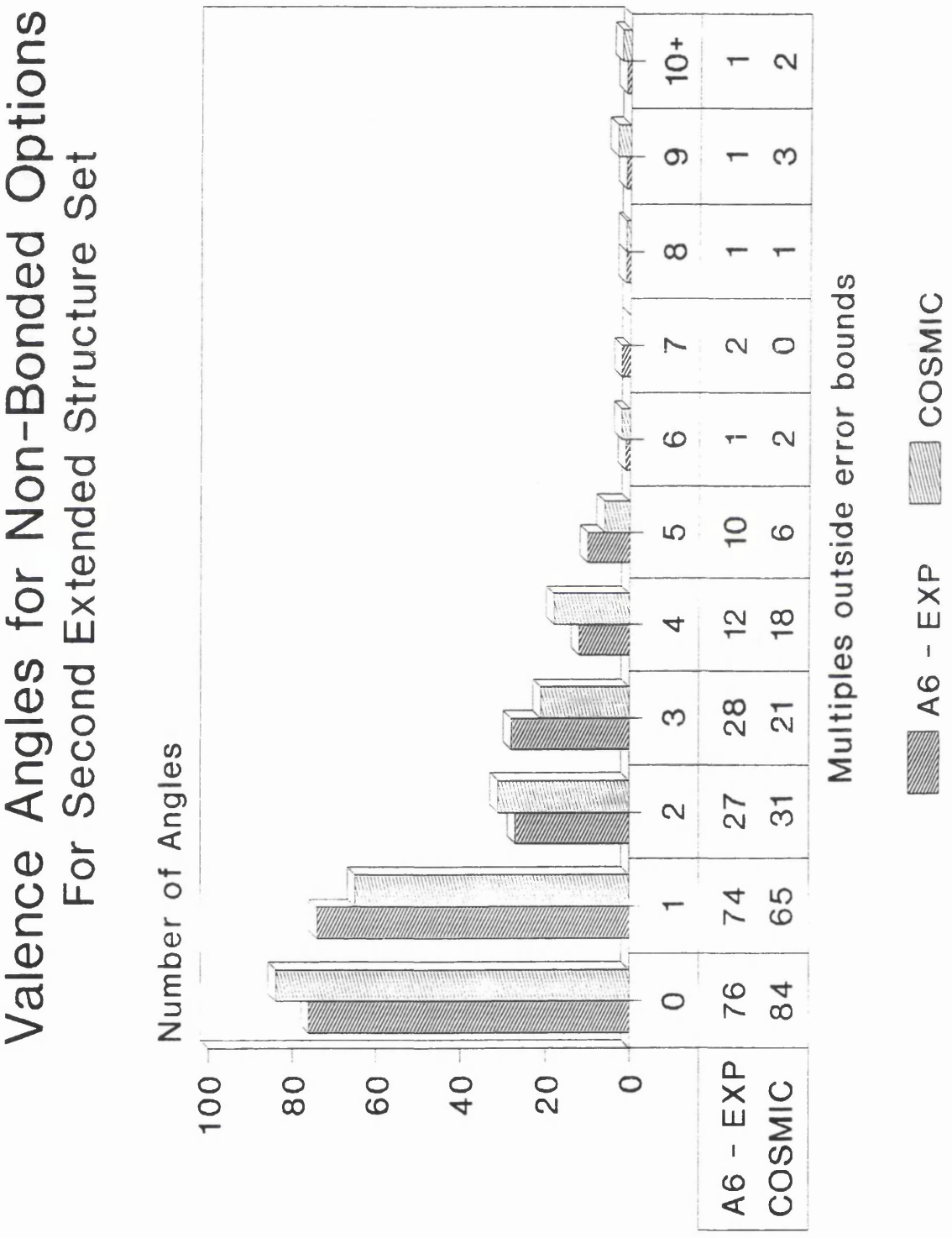


Figure 7.36 Torsion Angle Residual Factors for Non-Bonded's, 2nd Extended

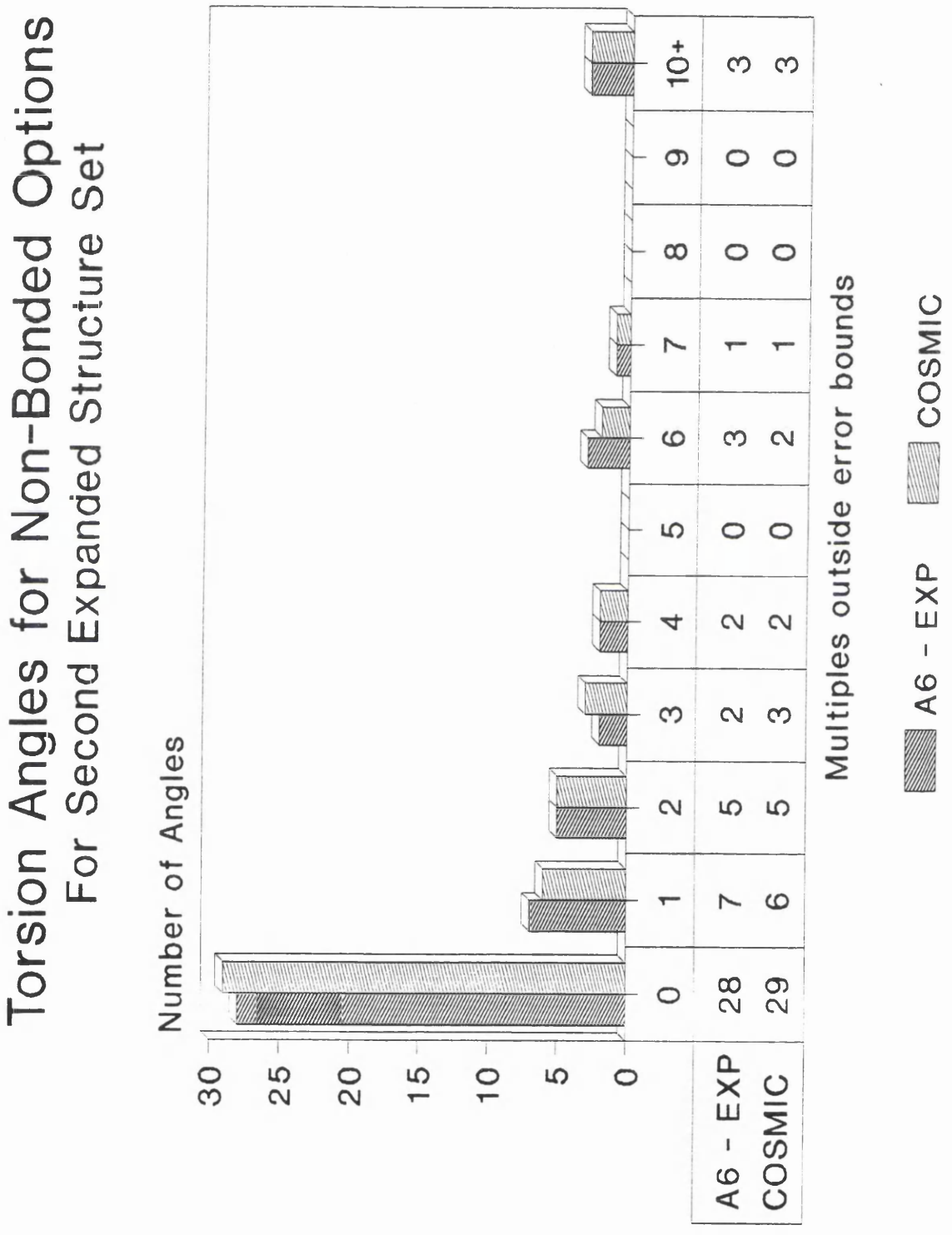
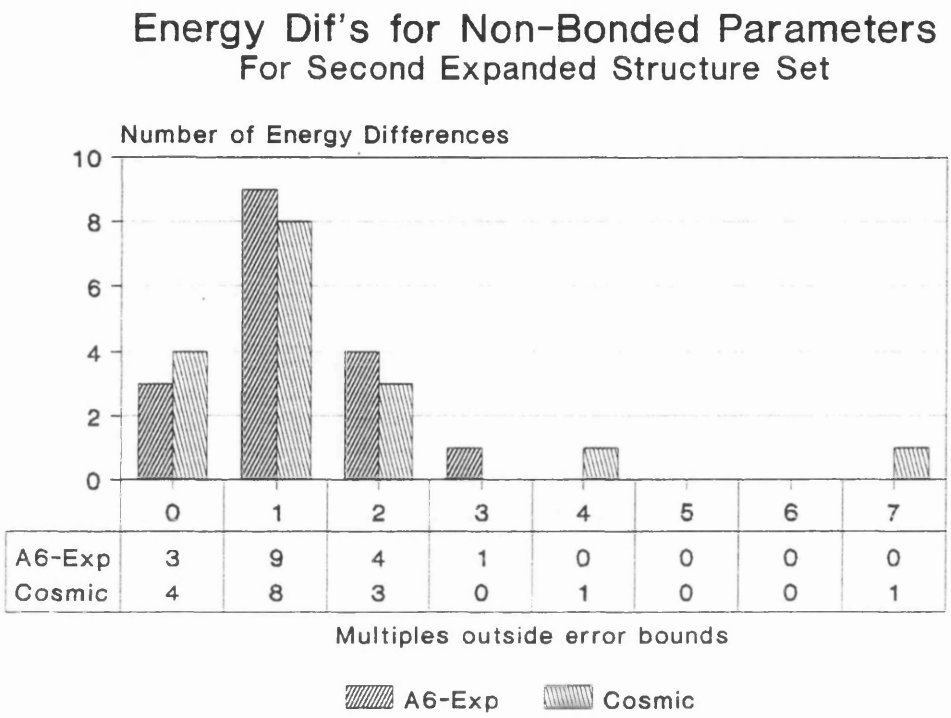
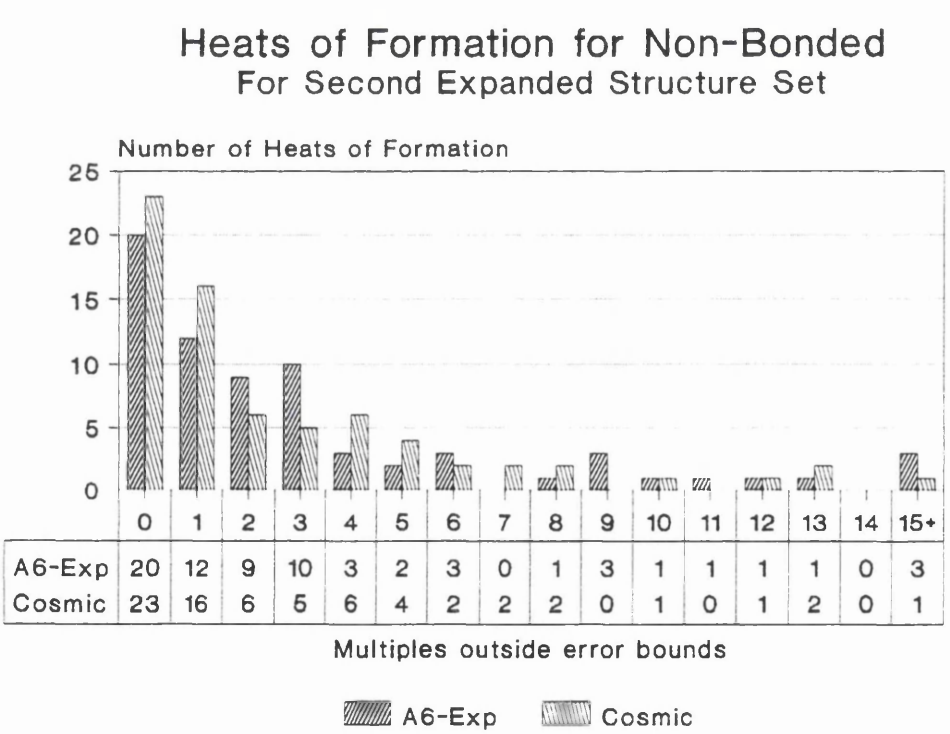


Figure 7.37 Isomer Energy Residual Factors for Non-Bonded's, 2nd Extended

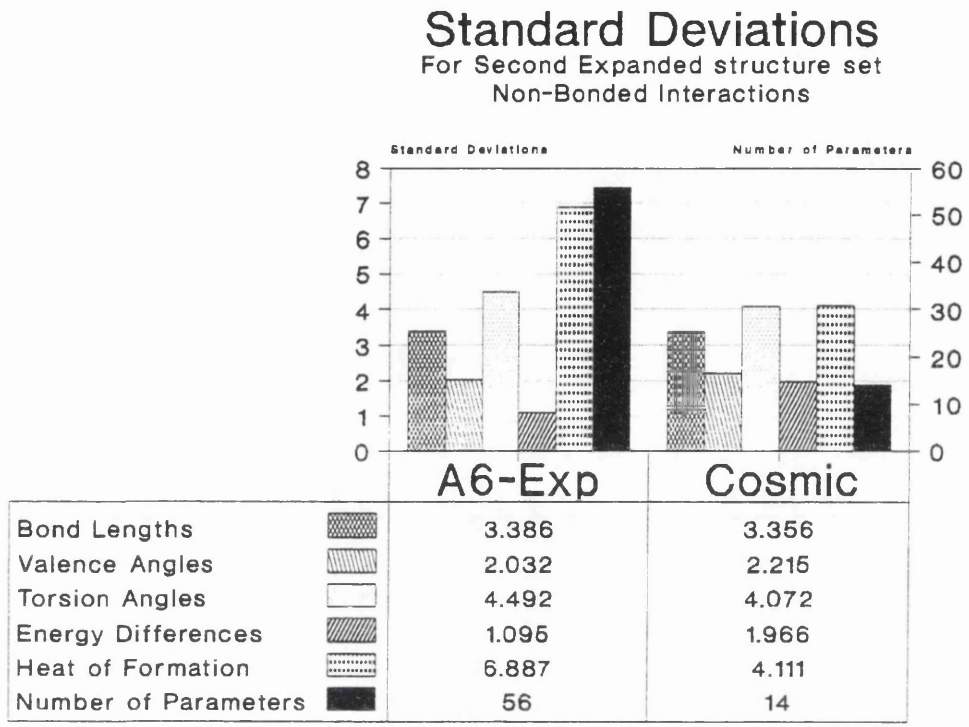


	0	1	2	3	4	5	6	7
A6-Exp	3	9	4	1	0	0	0	0
Cosmic	4	8	3	0	1	0	0	1

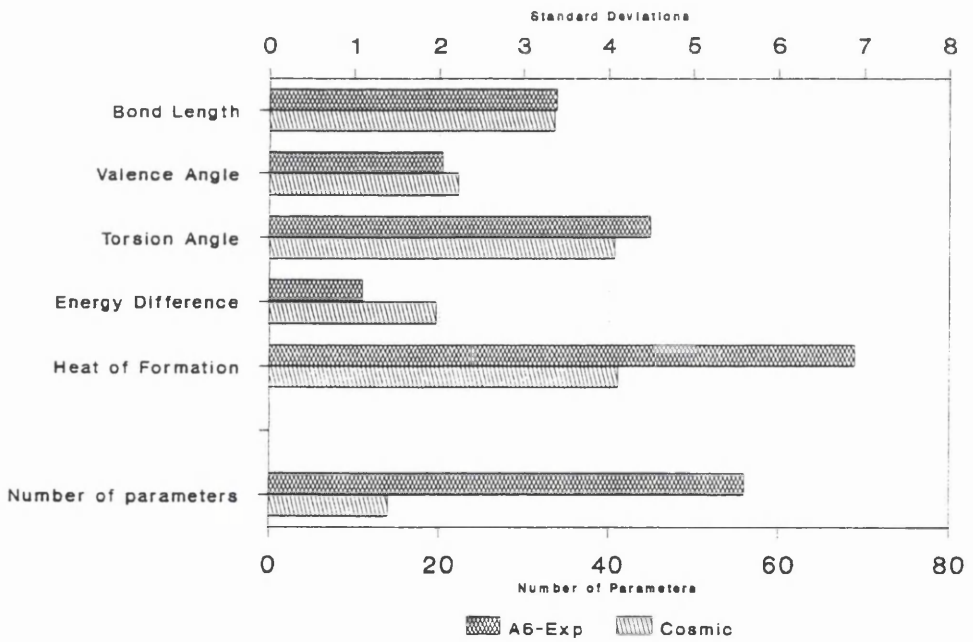


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
A6-Exp	20	12	9	10	3	2	3	0	1	3	1	1	1	1	0	3
Cosmic	23	16	6	5	6	4	2	2	2	0	1	0	1	2	0	1

Figure 7.38 Standard Deviations for Non-Bonded Options, 2nd Extended



	A6-Exp	Cosmic
Bond Lengths	3.386	3.356
Valence Angles	2.032	2.215
Torsion Angles	4.492	4.072
Energy Differences	1.095	1.966
Heat of Formation	6.887	4.111
Number of Parameters	56	14



Heat of Formation

Unlike the previous cases the heat of formation residual factor for the reduced parameter option is lower than that for the full case. Though the difference is not as drastic as in the angle parameter options. Why this should be is not clear, it may be that the low weighting means that the heat of formation residual factor has lost out to the highly weighted energy difference residual factor. That is that the non-bonded parameters have 'deformed' to produce a decrease in the energy difference residual factor that has more that made up for a resultant increase in the heat of formation residual factor.

As before it is useful to examine the number of parameters that would be required for both these approaches if they were used in the determination of a more expansive force field. For these cases the numbers of parameters are given in table 7.6.

Table 7.6 Comparison of Required Parameter for Non-Bonded Options

A6-EXP	≈1600
Cosmic	80

The deduction in this case must be that, given the closeness of the two methods in all areas, that the optimisability of the 'COSMIC' case makes it the optimum choice.

It should be noted that the non-bonded parameters are similar in their optimisation to the bond stretching force constants. The similarity lies in the fact that, for the majority of the structures, the non-bonded interaction is going to be small. An extra problem with the non-bonded interactions is that, for the electron density structures, there are very few non-bonded distances given. This means that the effect of non-bonding attraction or repulsion between atoms is only measurable by its effect on secondary residual factors such as torsion angles and valence angles. Because of this lack of a direct link between the interaction and the experimental properties the non-bonded parameters are also susceptible to deformation away from their normal perceived values when optimised using OPTPARAM.

7.4.4 Torsional Barriers

As explained earlier there are problems with the torsion angles in that they can adopt several low energy positions. If a molecule manages to obtain a minimum energy configuration different to that which is being compared then the result will be an unreasonably high residual factor the size of which will overwhelm the optimisation of the relevant parameter for those affected torsion angles that are still in the correct configuration.

To overcome this problem OPTPARAM 'fixes' the torsion angles to stay within 5 times the experimental error. Unfortunately this had not been implemented when the following comparisons were made.

Considering the number of parameters that would be required for a 'full' implementation the main aim of this investigation must be to determine if the reduced option is a credible alternative, as, without a valid reduced option the gains made in using fewer parameters in other options will be totally outweighed.

7.4.4.1 Alkene Structure Set

In the Alkene case the options compared were the case where the barriers are determined by just the central two atoms and the case taken from the White Bovill Force Field (WBFF) where the possible torsion angles are divided up into groups depending upon both the central atoms and on the atoms involved. The reason for not using the full case, where all possible combination of atoms in torsion bonds have different parameters, is that there is not enough experimental data in the alkene structure set to allow so many parameters to be optimised to any degree of accuracy.

Because of this the difference in the number of parameters between the cases studied is small and figures 7.39 and 7.40 show that there is little difference between the two cases. These graphs show the same trends as most of the other cases, i.e. the option with more parameters tends to produce lower standard deviations. In this case there is only one residual factor that is lower for the 'central two' case.

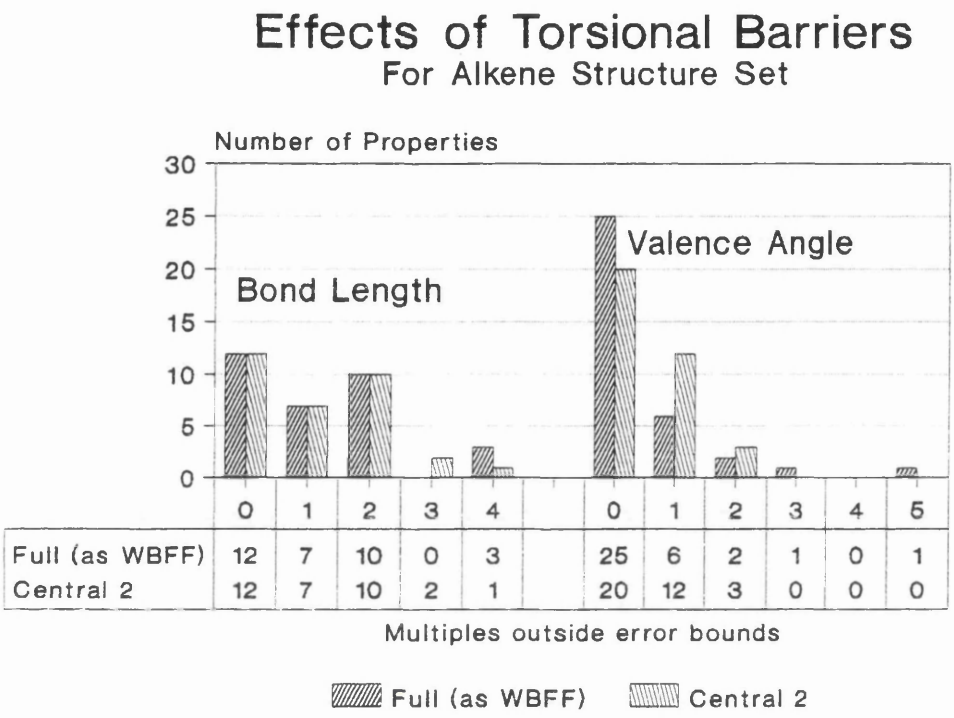
7.4.4.2 Expanded Structure Set

Because the Alkene comparison was between options that used roughly the same number of parameters a different approach was decided upon for the Expanded Structure Set.

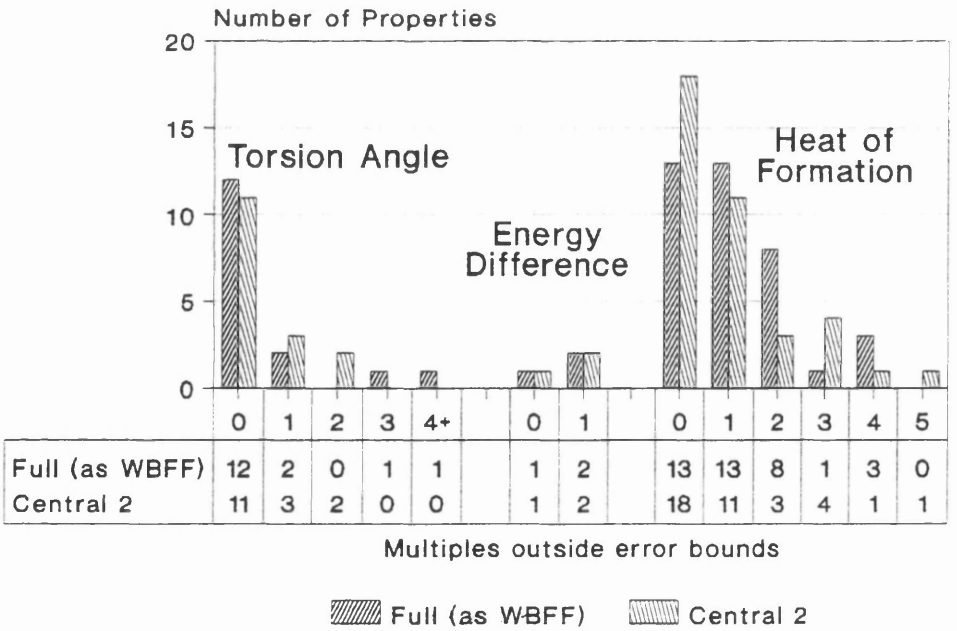
This approach involved ignoring one of the main points made so far, this being the need to make sure that there was a suitable depth of experimental data for optimisation. The notion here is that if the 'central two' option, where there is enough experimental data, could be shown to produce results close to an option that used more parameters but did not have enough data to produce reliable parameters then the use of the 'central two' option in the final optimisation would be justified.

In other words, if an optimisation is done with the full implementation then it is likely that the structure set will be deficient and so the optimisation results will be better than would be expected with a large, more balanced, structure set. So if the reduced option, on optimisation, is found to be relatively close to such an 'optimistic' optimisation, then it can be said that the results would be even closer for the 'global' structure set and so thus the reduced option could be validly used.

Figure 7.39 Residual Factors for Torsion Options, Alkene Study

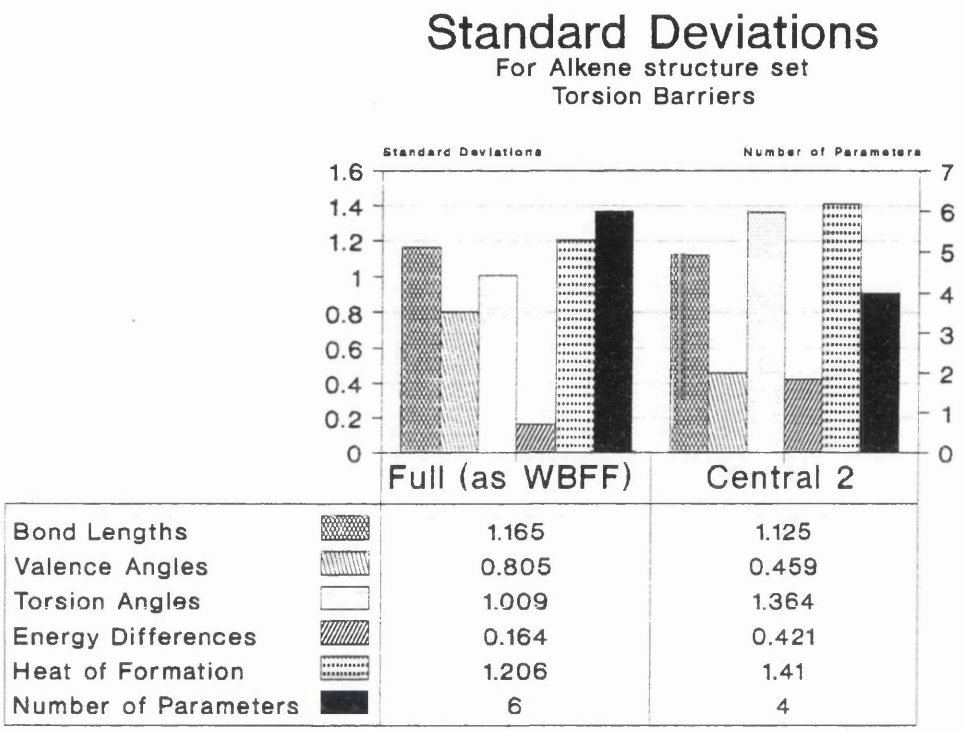


	Bond Length					Valence Angle					
	0	1	2	3	4	0	1	2	3	4	5
Full (as WBFF)	12	7	10	0	3	25	6	2	1	0	1
Central 2	12	7	10	2	1	20	12	3	0	0	0

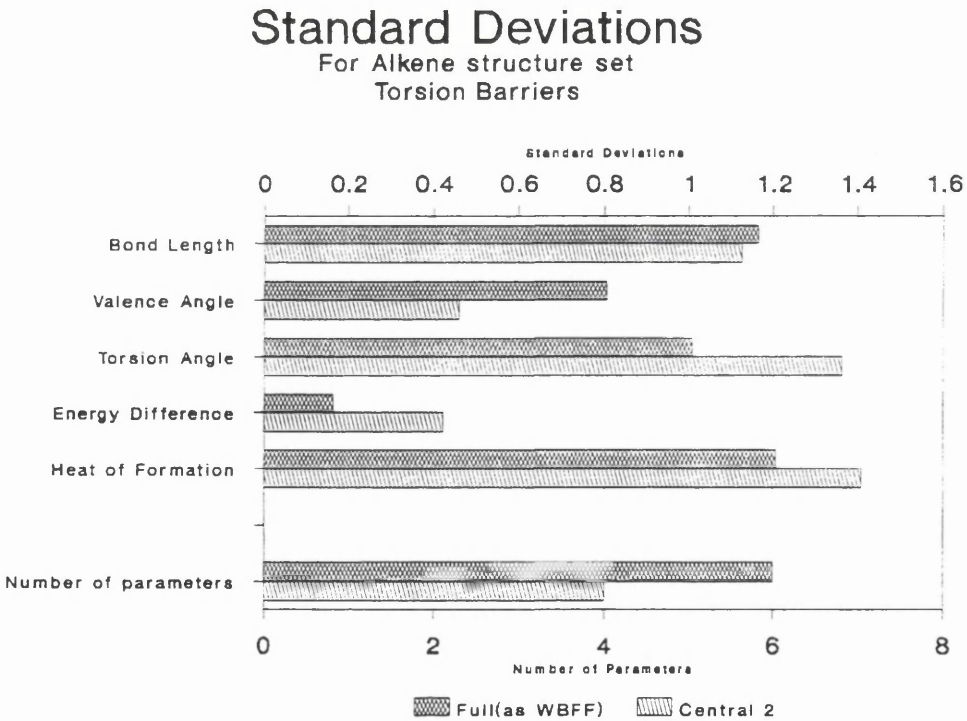


	Torsion Angle					Energy Difference		Heat of Formation					
	0	1	2	3	4+	0	1	0	1	2	3	4	5
Full (as WBFF)	12	2	0	1	1	1	2	13	13	8	1	3	0
Central 2	11	3	2	0	0	1	2	18	11	3	4	1	1

Figure 7.40 Standard Deviations for Torsion Options, Alkene Study



	Full (as WBFF)	Central 2
Bond Lengths	1.165	1.125
Valence Angles	0.805	0.459
Torsion Angles	1.009	1.364
Energy Differences	0.164	0.421
Heat of Formation	1.206	1.41
Number of Parameters	6	4



The full parameter option used in this test is a slightly reduced version of the maximum possible case. In this case there are individual parameters for all possible torsion angles except that there is a single torsion barrier for the $C_{sp2}=C_{sp2}$ bond. The reasoning for this is that there is little, to no, rotation about such bonds and so a different parameter for all of these cases would be excessive. This option is designated as 'full(trial)' in the following results. With this option there are 65 parameters to be optimised, the Second Expanded Structure set only contains information of 46 torsion angles so it can be seen that for this case the likelihood of the production of structure set related artefacts for this parameter selection option is high.

When the comparison of the torsion angle types using the standard weightings of structure 5, energy difference 20, heat of formation 1 was investigated it was found that the optimisation for the 'full(trial)' option resulted in more of the torsion angles ending up in the incorrect minima as compared to the 'central 2' case. This problem was not noticed at the time and it makes it impossible to compare the two approaches evenly with these results. Because of this, earlier results were investigated and two comparable optimisations were found. Unlike the later optimisations the weighting in this case was structure 5, energy difference 1, heat of formation 1.

Figures 7.41 to 7.45 show the results of the optimisation on these two cases. The most remarkable thing about these results is the closeness of all the residual factors with neither of the methods showing any advantage over the other. To emphasise this point investigation shows that the standard deviations differ by an average of only 1.5%. Given the proximity of these results and the large difference in the number of parameters it is a simple task to say that, on this evidence, for the torsional barriers the 'central 2' option is the optimum one for use in the final force field optimisation.

Figure 7.41 Torsion Angle Residual Factors for Torsion Barrier Options

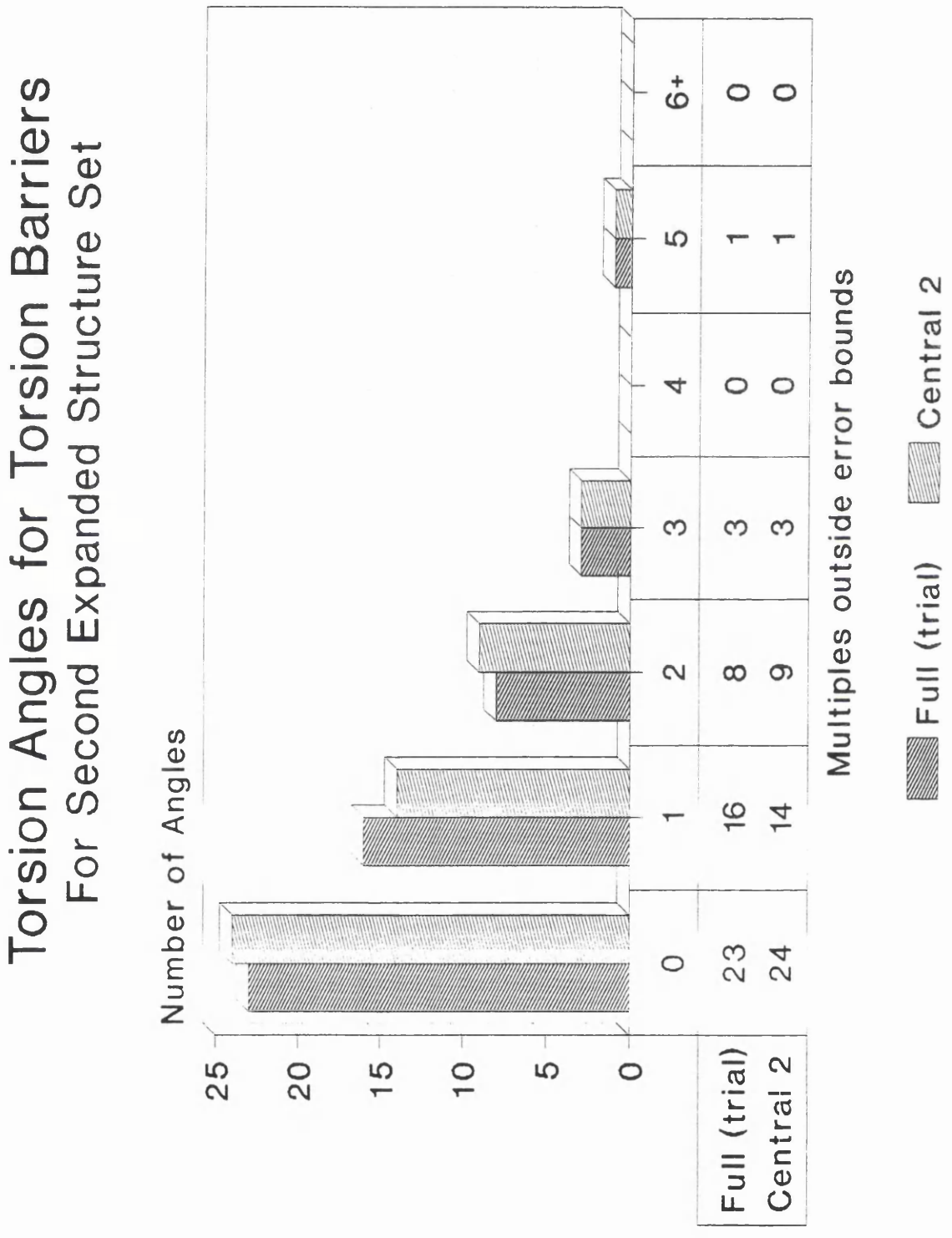


Figure 7.42 Distance Residual Factors for Torsion Barrier Options

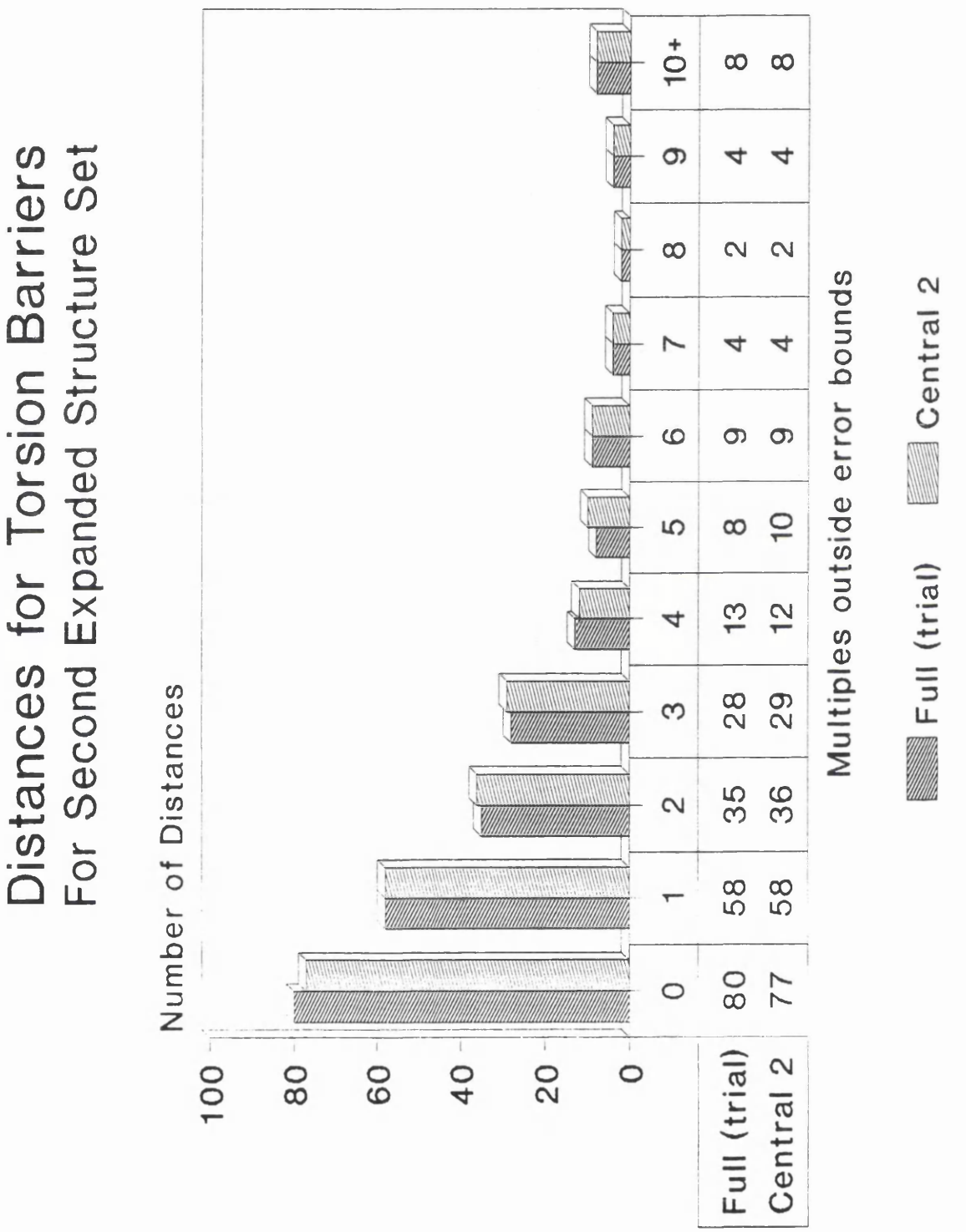


Figure 7.43 Valence Angle Residual Factors for Torsion Barrier Options

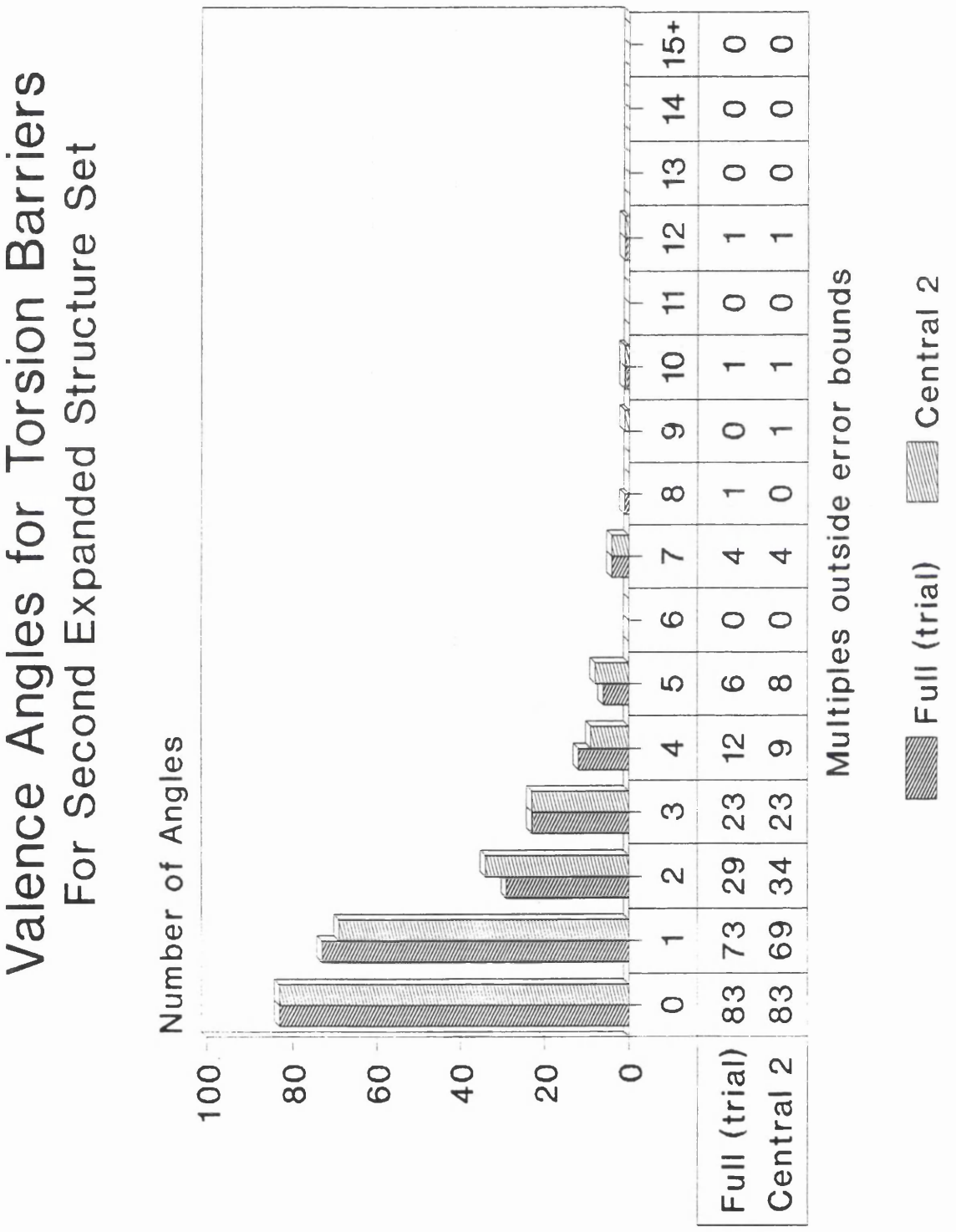
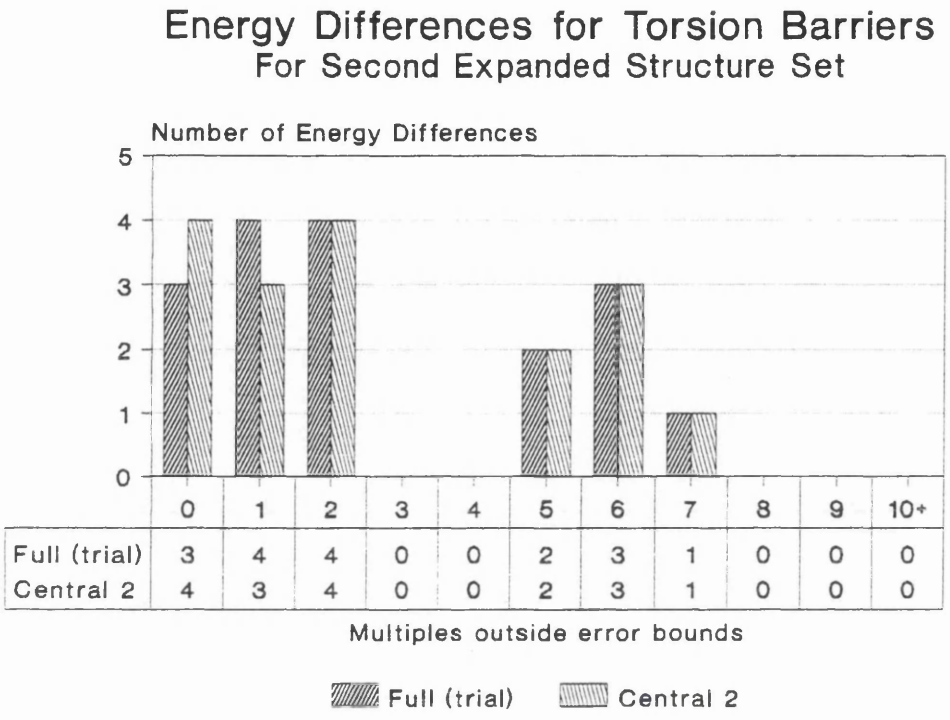


Figure 7.44 Isomeric Energy Residual Factors for Torsion Barrier Options



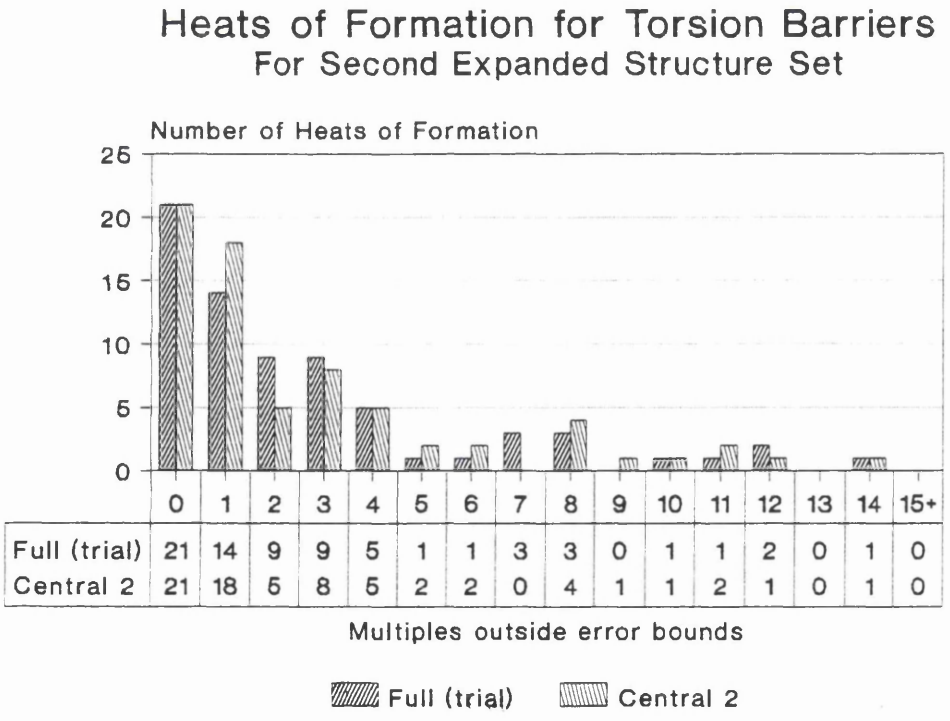
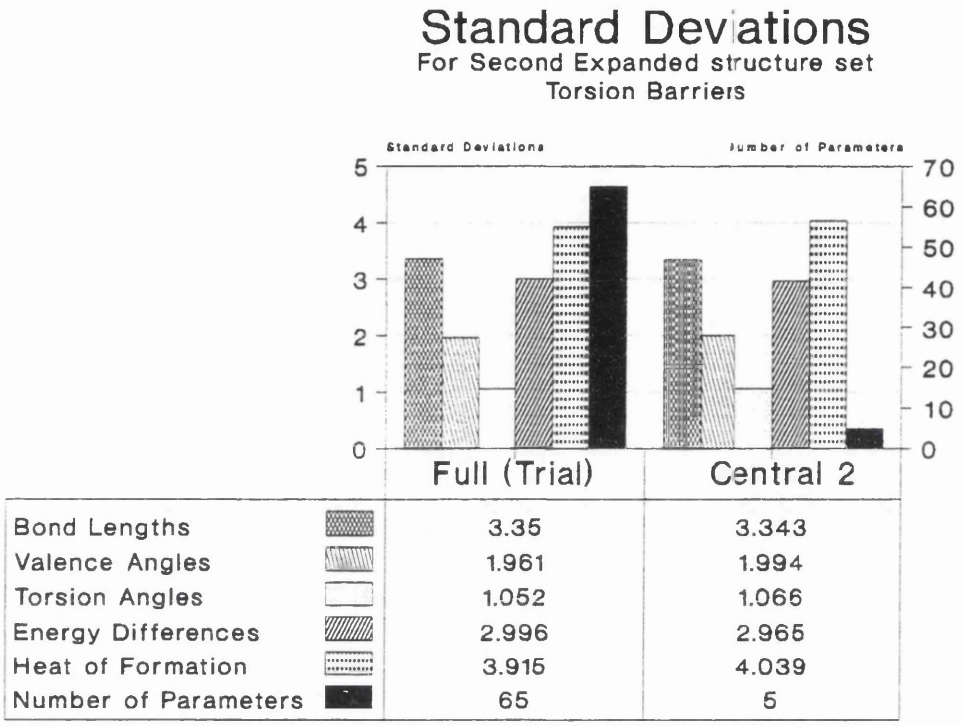
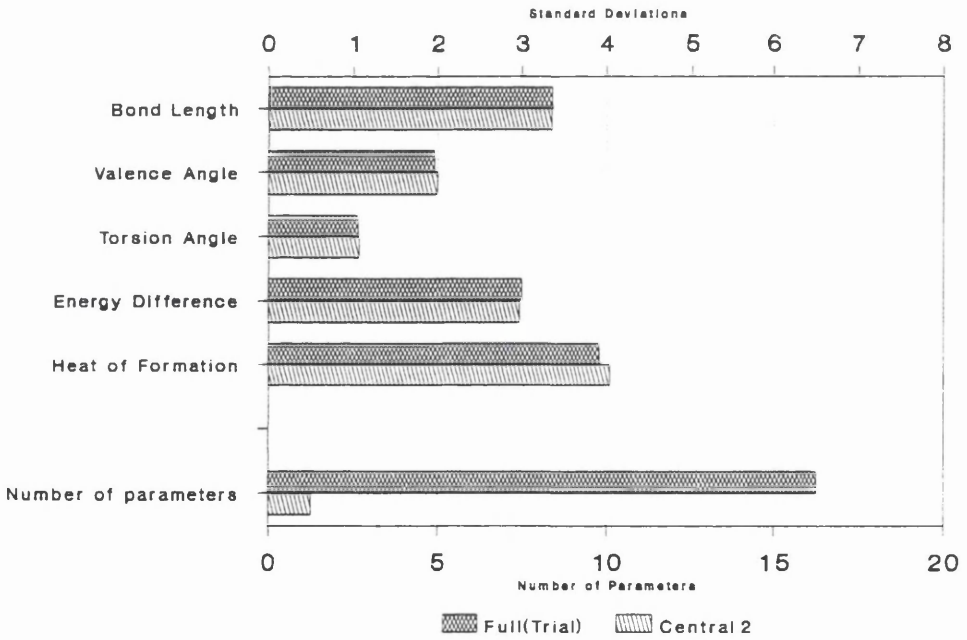


Figure 7.45 Standard Deviations for Torsion Barrier Options



	Full (Trial)	Central 2
Bond Lengths	3.35	3.343
Valence Angles	1.961	1.994
Torsion Angles	1.052	1.066
Energy Differences	2.996	2.965
Heat of Formation	3.915	4.039
Number of Parameters	65	5



7.5 Final Optimisation

As discussed in the above sections the following force field configuration was chosen as the final configuration for use with the Second Expanded Structure Set.

Bond Length	Full
Bond Stretching	Calculated
Non-Bonded	COSMIC
Valence Angle	Calculated
Angle Bend	Calculated
Torsion Angle	Central Two

An optimisation was set up with this configuration and left to optimise for 100 force field loops, which took 226 hours. Figure 7.46 shows the alteration in the residual factors for this run. It can be seen that, ignoring the small disturbances, the residual factors are very level by the end of the run.

This can be shown quantitatively by investigating the gradients of the curve for the total residual factor over selected ranges. It is found that over the first 50 parameter optimisations a least squares fit for a straight line gives a gradient of -0.1117, this can be compared with the last 1500 parameter optimisations, where the gradient is down to -0.0003, or 0.27%, of its initial value. In other words it is taking over 6 full force field loops at the end to produce the same drop in total residual factor as a single parameter optimisation at the beginning.

Unfortunately, due to limited disk space at the time of the final optimisation, the parameter blocks were not saved at the end of each force field loop, so it is impossible to see if movement in the actual parameters had reduced to a similar minimal level, but after such a large number of iterations it seems likely.

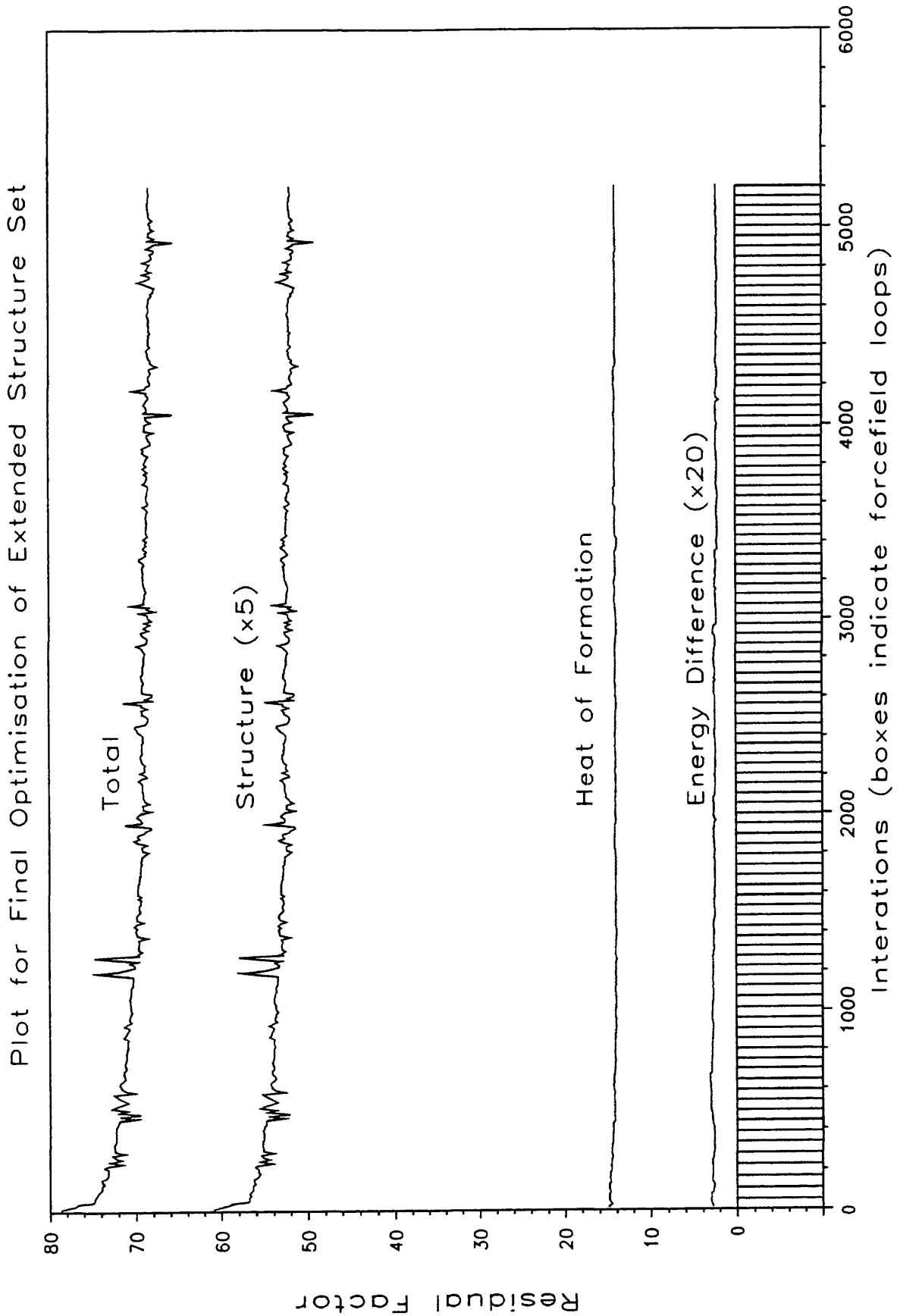
In the following sections the results of this optimisation will be given. First the comparison of the results with the experimental data in the structure set will be investigated, then the actual parameters will be listed and their generality examined.

7.5.1 Overall Results

The following section is an extract from the end of the final COMPARE file. It gives a summary of the overall results for the optimisation.

The first section compares the average differences between the experimental values and the calculated values with the average experimental errors. It shows that for the bond lengths, valence angles and energy differences the average experimental to calculated difference is very close to the average experimental cases. For the

Figure 7.46 Residual Factor Plot for Final Optimisation



torsion angles the average experimental to calculated difference is roughly double the average experimental error. The worst case is the heats of formation, almost certainly because the artificial nature of the geometric enthalpy increments will not work effectively with some of the structures.

The results in all these cases will be examined later on in this section.

Extract From Final COMPARE File

```
AVERAGE BOND LENGTH DIFFERENCE         :-   0.0153    AVERAGE ERROR  :-   0.0244

AVERAGE BOND ANGLE DIFFERENCE          :-   1.6819    AVERAGE ERROR  :-   1.1635

AVERAGE TORSION ANGLE DIFFERENCE       :-   7.1215    AVERAGE ERROR  :-   3.7882

AVERAGE CONFORMER DIFFERENCE           :-   0.4446    AVERAGE ERROR  :-   0.3084

AVERAGE HEAT OF FORMATION DIFFERENCE   :-   2.5219    AVERAGE ERROR  :-   0.6726

AVERAGE OF SUM 1st    :-   0.0166

NUMBER OF PARAMETERS USED :-50

 ITERATIONS 100, SWEIGHT   5.0, EWIEGHT  20.0, HWEIGHT   1.0

 FILES USED FOR THIS OPTIMISATION

 BOND LENGTH , FULL , NUMBER 1
 BOND STRENGTH , CALCULATED , NUMBER 1
 NON BONDED , COSMIC TYPE , NUMBER 1
 ANGLE VALUES , CALCULATED , NUMBER 1
 ANGLE BEND , CALCULATED , NUMBER 1
 TORSION , CENTRAL TWO , NUMBER 1
 OUT OF PLANE BENDING , IMPLEMENTED , NUMBER 1

 STRUCTURE          :      10.36323
 ENERGY             :       0.11694
 HEAT OF FORMATION  :      14.11788

 TOTAL              :      68.27274
```

7.5.1.1 Structure Results

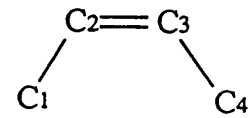
As mentioned before, to save disk space, the COMPARE file only listed those structure elements where the calculated value was outside the error of the experimental value. This option was introduced during the first trials of a mixed electron, X-ray & neutron diffraction structure set (chapter 8). In this case a COMPARE file which contained all the structural properties took up 1200KBytes, whereas the same COMPARE file with only those structural properties outside the relevant experimental error given usually took up 360KBytes of disk space. In

retrospect this option should have been disabled for the final optimisation, unfortunately due to timing problems involved in getting an optimisation with a significant number of iterations, it was overlooked.

Because of this specific information is only available on those geometric properties that, for the final force field, were outside their experimental error bounds. The following section lists the structures for which experimental geometric properties were used and for each the experimental and calculated values, when available, are given. Where the calculated value from the final optimisation was not saved this is noted using '-----' where the result would have be given.

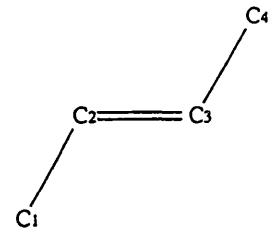
Cis But-2-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.506(2)	-----	-----
C_2=C_3	1.346(3)	-----	-----
C_1-C_2=C_3	125.4(4)	-----	-----



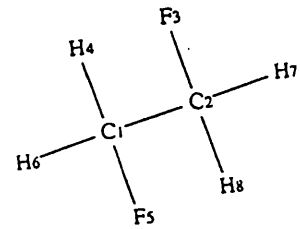
Trans But-2-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.508(2)	-----	-----
C_2=C_3	1.347(3)	-----	-----
C_1-C_2=C_3	123.8(4)	124.8	1.0



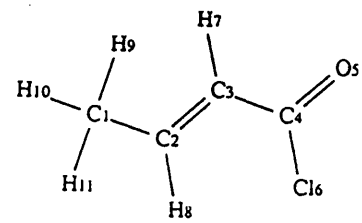
Anti 1,2-Difluoro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-H_4	1.103(4)	-----	-----
C_1-C_2	1.503(3)	1.520	0.017
C_1-F_5	1.389(2)	1.345	0.045
H_4-C_1-H_6	108.7(16)	-----	-----
C_1-C_2-F_3	110.3(1)	109.9	0.4
C_1-C_2-H_7	110.3(10)	112.3	2.0



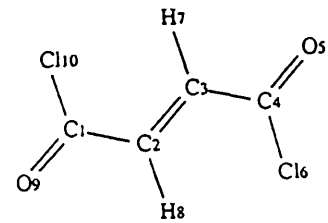
Anti Trans-2-Butenoyl Chloride

Molecular Property	Exptl Value	Calc Value	Outside Error
O_5=C_1-C_2=C_3	0.0(50)	-----	-----



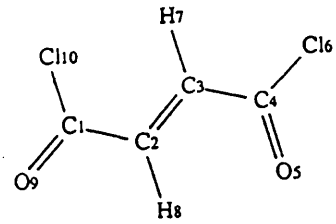
Anti-Anti Fumaryl Chloride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-H_8	1.089(25)	-----	-----
C_1=O_9	1.191(2)	1.198	0.005
C_2=C_3	1.334(5)	1.346	0.012
C_1-C_2	1.488(3)	1.470	0.018
C_4-Cl_6	1.783(2)	1.750	0.033
C_1-C_2=C_3	125.2(6)	-----	-----
O_9=C_1-C_2	125.5(4)	126.1	0.6
Cl_6-C_4-C_3	114.6(3)	115.6	1.0



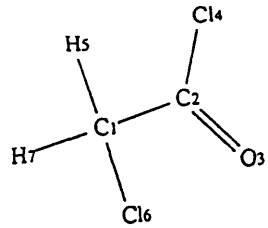
Anti-Syn Fumaryl Chloride

Molecular Property	Exptl Value	Calc Value	Outside Error
Cl_6-C_4-C_3=C_2	155.3(55)	180.0	24.7



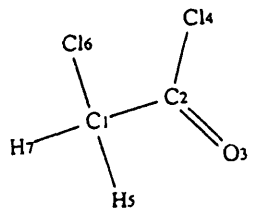
Gauche Chloro Acetyl Chloride

Molecular Property	Exptl Value	Calc Value	Outside Error
Cl_4-C2-C_1-Cl_6	116.4(77)	124.2	7.8



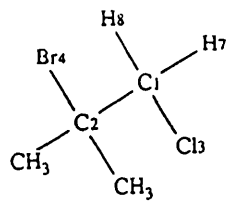
Anti Chloro Acetyl Chloride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-Cl_6	1.782(18)	-----	-----
C_1-H_5	1.062(30)	1.094	0.032
C_2=O_3	1.182(4)	1.198	0.016
C_1-C_2	1.521(9)	1.487	0.034
C_2-Cl_4	1.772(16)	1.749	0.023
C_2-C_1-Cl_6	112.9(17)	-----	-----
C_1-C_2=O_3	126.9(9)	125.1	1.8
C_1-C_2-Cl_4	110.0(7)	111.1	1.1
H_5-C_1-H_7	109.5(20)	106.2	3.3
Cl_6-C_1-C_2-Cl_4	0.0(50)	-----	-----



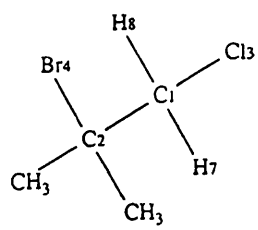
Anti 2-Bromo-3-Chloro-2-Methyl Propane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-Cl_3	1.798(5)	1.775	0.023
C_2-Br_4	1.989(8)	1.978	0.011
CL_3-C_1-C_2	115.5(14)	111.4	4.1
Br_4-C_2-C_1-Cl_3	159.0(30)	174.1	15.1



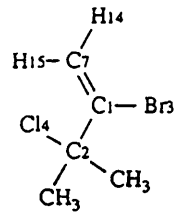
Gauche 2-Bromo-3-Chloro-2-Methyl Propane

Molecular Property	Exptl Value	Calc Value	Outside Error
Br_4-C_2-C_1-Cl_3	54.0(40)	61.9	7.9



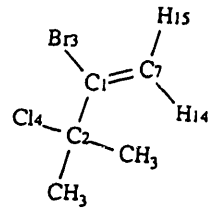
Anti 2-Bromo-3-Chloro Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.491(16)	-----	-----
C_1-Br_3	1.911(11)	-----	-----
C_1=C_7	1.360(14)	1.331	0.029
C_2-Cl_4	1.803(9)	1.778	0.025
C_7=C_1-Br_3	121.3(30)	-----	-----
C_1-C_2-Cl_4	113.0(26)	-----	-----
C_2-C_1=C_7	128.4(10)	123.9	4.5



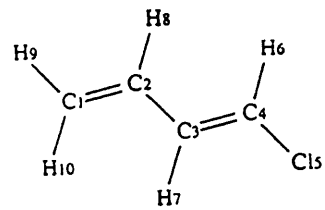
Gauche 2-Bromo-3-Chloro Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
Br_3-C_2-C_1-Cl_4	70.5(35)	-----	-----



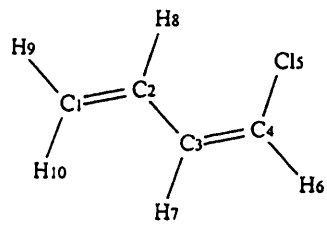
Trans 1-Chloro-1,3-Butadiene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_3=C_4	1.340(2)	-----	-----
C_1=C_2	1.344(2)	1.341	0.003
C_2-C_3	1.461(3)	1.469	0.008
C_4-Cl_5	1.728(2)	1.747	0.019
C_2-C_3=C_4	122.5(4)	123.1	0.006
C_1=C_2-C_3	123.3(3)	123.9	0.006
C_3=C_4-Cl_5	122.5(2)	122.8	0.003



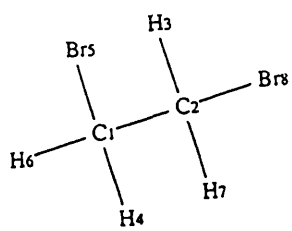
Cis 1-Chloro-1,3-Butadiene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_3=C_4	1.342(2)	1.334	0.008
C_1=C_2	1.345(2)	1.340	0.005
C_2-C_3	1.466(3)	1.472	0.006
C_4-Cl_5	1.730(2)	1.745	0.015
C_2-C_3=C_4	125.5(2)	124.9	0.6
C_1=C_2-C_3	122.3(4)	123.8	1.5
C_3=C_4-Cl_5	123.9(2)	121.5	2.4



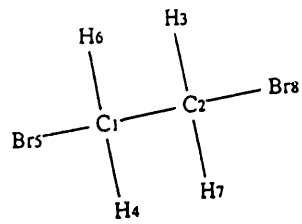
Gauche 1,2-Dibromo Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
Br_5-C_1-C_2-Br_8	73.0(46)	66.1	6.9



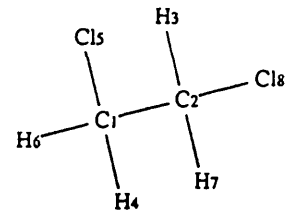
Anti 1,2-Dibromo Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.506(7)	-----	-----
C_1-Br_5	1.950(3)	-----	-----
C_1-H_4	1.108(8)	1.100	0.008
C_1-C_2-H_3	110.0(11)	-----	-----
C_1-C_2-Br_8	109.5(4)	110.3	0.8



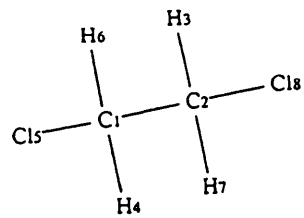
Gauche 1,2-Dichloro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
Cl_5-C_1-C_2-Cl_8	72.4(31)	65.8	6.6



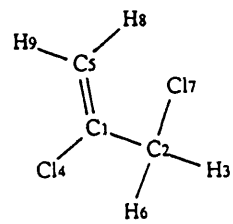
Anti 1,2-Dichloro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.510(6)	-----	-----
C_1-Cl_5	1.788(2)	1.771	0.017
C_1-H_4	1.081(8)	1.099	0.018
C_1-C_2-Br_8	109.2(3)	109.9	0.7
C_1-C_2-H_3	110.9(5)	110.1	0.8



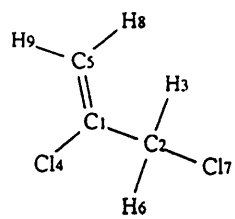
Anti 2,3-Dichloro-1-Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_5	1.334(9)	-----	-----
C_1-C_2	1.504(13)	1.487	0.017
C_2-C_1=C_5	127.6(11)	124.5	3.1
C_2-C_1-Cl_4	110.2(10)	114.8	4.6
C_1-C_2-Cl_7	113.1(12)	114.3	1.2



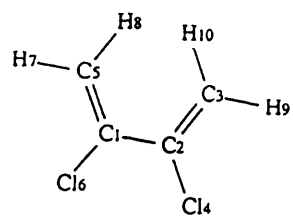
Gauche 2,3-Dichloro-1-Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_5=C_1-C_2-Cl_7	108.9(34)	113.1	4.2



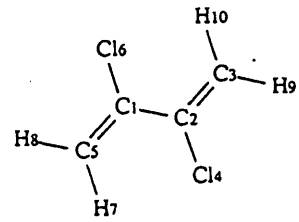
Gauche 2,3-Dichloro-1,3-Butadiene

Molecular Property	Exptl Value	Calc Value	Outside Error
Cl_6-C_1-C_2-Cl_4	52.3(97)	3.7	48.6



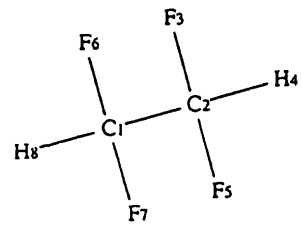
Anti 2,3-Dichloro-1,3-Butadiene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.463(8)	-----	-----
C_3-H_9	1.101(11)	1.086	0.015
C_2=C_3	1.341(4)	1.330	0.011
C_2-Cl_4	1.744(3)	1.751	0.007
C_1-C_2=C_3	126.0(4)	123.8	2.2
C_1-C_2-Cl_4	115.8(5)	118.2	2.4
C_2=C_3-H_9	120.6(25)	124.0	3.4



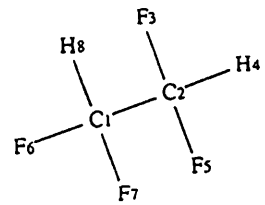
Anti 1,1,2,2-Tetrafluoro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.518(5)	-----	-----
C_1-H_8	1.098(6)	-----	-----
C_1-F_6	1.350(2)	1.345	0.005
C_1-C_2-F_3	108.2(3)	111.1	2.9
F_6-C_1-F_7	107.3(3)	105.7	1.6
C_1-C_2-H_4	110.3(10)	112.9	2.6



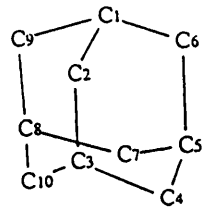
Gauche 1,1,2,2-Tetrafluoro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
H_8-C_1-C_2-H_4	78.0(2)	68.0	10.0



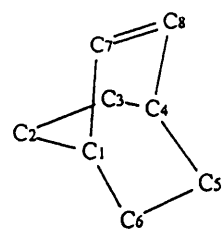
Adamantane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.534(4)	1.545	0.011
C_2-C_3-C_4	109.2(5)	-----	-----
C_1-C_2-C_3	110.0(5)	109.2	0.8



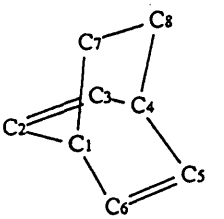
Bicyclo [2.2.2] Oct-2-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_6-C_1••C_4-C_3	121.2(20)	-----	-----
C_1-C_7=C_8-C_4	0.0(20)	-----	-----



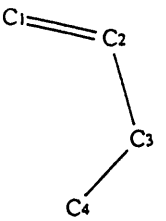
Bicyclo [2.2] Octa-2,5-Diene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_6-C_1••C_4-C_3	123.4(20)	-----	-----
C_1-C_7=C_8-C_4	0.0(20)	-----	-----



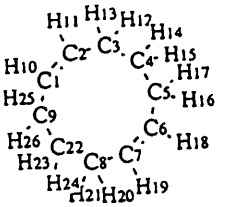
Cis But-1-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C1=C_2-C_3	126.7(5)	-----	-----
C_2-C_3-C_4	114.8(5)	117.4	2.6



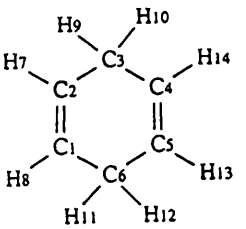
Cyclodeca-1,6-diene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_4-H_{14}	1.112(4)	1.108	0.004
C_1=C_2	1.326(4)	1.352	0.026
C_3-C_4	1.534(6)	1.545	0.011
C_2-C_3	1.506(6)	1.516	0.010
H_{14}-C_4-H_{15}	105.6(10)	-----	-----
C_1=C_2-C_3	128.2(3)	129.7	1.5
C_2-C_3-C_4	112.8(3)	112.2	0.6
C_3-C_4-C_5	114.1(5)	113.3	0.8
H_{12}-C_3-H_{13}	116.6(10)	118.3	1.7



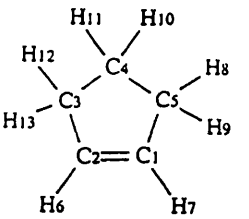
Cyclohexa-1,4-diene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2	1.334(2)	-----	-----
C_6-H_{12}	1.114(3)	1.101	0.013
C_1-H_8	1.103(3)	1.096	0.007
C_1-C_6	1.496(2)	1.499	0.003
C_2-C_1-H_8	123.4(27)	-----	-----
C_2-C_3-C_4	113.3(3)	114.2	0.9
C_3-C_4=C_5	123.4(2)	122.9	0.5
C_1-C_6-H_{11}	110.0(4)	109.0	1.0
H_{11}-C_6-H_{12}	103.0(20)	106.3	3.3
C_6-C_1-H_8	113.2(27)	115.9	2.7



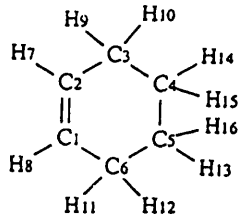
Cyclopentene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_5••C_3-C_4	156.7(20)	-----	-----



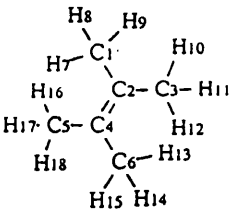
Cyclohexene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-C_3	1.504(6)	-----	-----
C_1-H_8	1.093(15)	-----	-----
C_1=C_2	1.335(3)	1.340	0.005
C_3-C_4	1.515(20)	1.538	0.023
C_1=C_2-C_3	123.5(5)	-----	-----
C_2-C_3-C_4	112.1(5)	-----	-----
C_3-C_4-C_5	110.0(5)	-----	-----
C_1=C_2-C_3-C_4	15.2(20)	-----	-----
C_2-C_3-C_4-C_5	-44.9(20)	-----	-----
C_3-C_4-C_5-C_6	60.2(20)	-----	-----



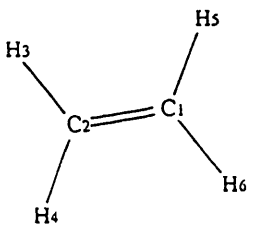
2.3-Dimethyl But-2-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2=C_4	1.353(4)	1.368	0.015
C_1-C_2	1.511(2)	1.523	0.012
C_1-C_2=C_4	123.9(5)	123.0	0.9



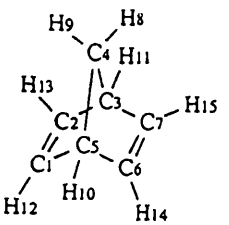
Ethylene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2	1.335(3)	-----	-----
C_1-H_3	1.090(3)	-----	-----
C_1=C_2-H_5	121.7(4)	123.0	1.3



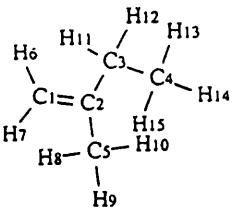
Norbornadiene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-C_3	1.535(7)	1.500	0.035
C_1=C_2	1.343(3)	1.330	0.013
C_3-C_4	1.544(4)	1.529	0.015
C_3-C_4-C_5	94.1(30)	-----	-----
C_2-C_3••C_5-C_6	115.6(20)	113.4	2.2



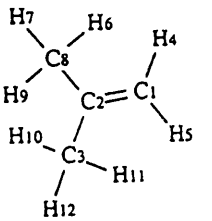
2-Methyl Skew But-1-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_5-C_2-C_3-C_4	72.7(25)	85.3	12.6



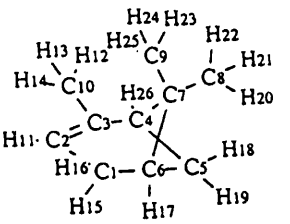
Isopropene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2-C_3	122.1(3)	-----	-----



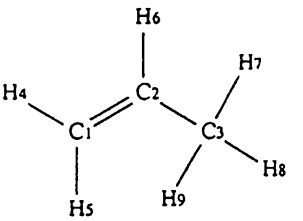
Pin-2-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2=C_3	1.340(10)	1.353	0.013
C_2-C_1-C_6	112.0(30)	-----	-----
C_2=C_3-C_4	118.0(30)	-----	-----
C_3=C_2-C_1	118.0(30)	-----	-----
C_2-C_3-C_{10}	126.0(30)	-----	-----
C_7-C_4••C_6-C_5	156.0(80)	-----	-----



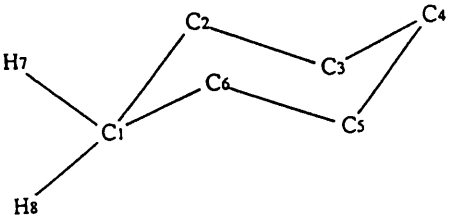
Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2	1.341(1)	-----	-----
C_2-C_3	1.504(1)	1.505	0.001
C_1=C_2-C_3	124.8(1)	124.7	0.1



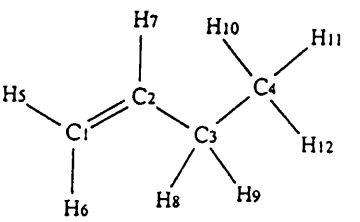
Cyclohexane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.528(4)	1.540	0.008
C_1-H_7	1.100(4)	1.108	0.008
C_1-C_2-C_3	111.3(5)	-----	-----
H_7-C_1-H_8	110.0(25)	105.7	4.3
C_1-C_2-C_3-C_4	55.2(5)	54.7	0.5



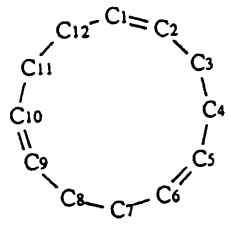
Skew But-1-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2-C_3	125.4(5)	124.5	0.9
C_2-C_3-C_4	112.1(5)	111.1	1.0



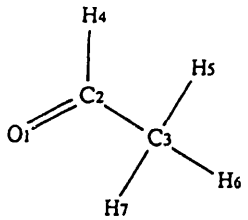
Trans,Trans,Trans Cyclododeca-1,5,9-triene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_3-C_4	1.540(10)	-----	-----
C_2-C_3	1.490(10)	1.512	0.022
C_1=C_2	1.320(10)	1.344	0.024
C_3-C_4-C_5	111.1(5)	-----	-----
C_4-C_5=C_6	124.1(5)	124.7	0.6
C_3-C_4-C_5=C_6	116.5(20)	-----	-----
C_2-C_3-C_4-C_5	63.4(20)	57.5	5.9
C_4-C_5=C_6-C_7	178.0(20)	-171.4	4.6



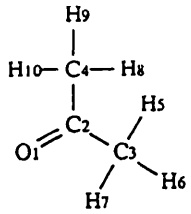
Acetaldehyde

Molecular Property	Exptl Value	Calc Value	Outside Error
C_3-H_4	1.086(23)	-----	-----
O_1=C_2	1.208(3)	1.200	0.008
C_2-C_3	1.514(5)	1.508	0.006
O_1••C_3	2.400(9)	2.414	0.014



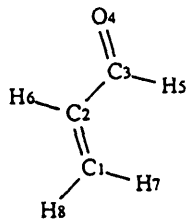
Acetone

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-C_3	1.517(3)	-----	-----
O_1=C_2	1.210(4)	1.204	0.006
C_3-H_5	1.091(3)	1.100	0.009
H_5-C_3-H_6	108.5(5)	106.4	2.1
C_3-C_2-C_4	116.0(3)	113.7	2.3



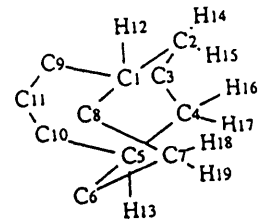
Acrolein

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-H_6	1.079(6)	1.096	0.017
C_3=O_4	1.209(2)	1.198	0.011
C_1=C_2	1.340(3)	1.344	0.004
C_2-C_3	1.481(4)	1.474	0.007
C_2-C_4=O_4	124.7(7)	-----	-----
C_1=C_2-H_6	121.6(40)	-----	-----
C_2-C_3-H_5	115.6(30)	-----	-----
C_1=C_2-C_3	119.8(10)	125.0	5.2



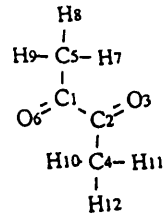
Manxane

Molecular Property	Exptl Value	Calc Value	Outside Error
H_{15}-H_{18}	2.200(40)	2.428	0.208



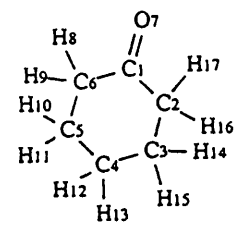
2,3 Butanedione

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_5	1.527(6)	-----	-----
C_2=O_3	1.214(2)	1.207	0.007
C_5-H_7	1.114(7)	1.099	0.015
C_1-C_2-C_4	116.3(3)	116.8	0.5
C_1-C_2=O_3	120.3(7)	117.3	3.0
C_1-C_5-H_7	108.1(16)	114.1	6.0



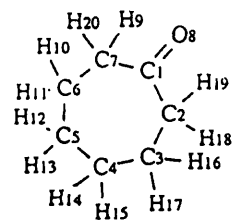
Cyclohexone

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=O_7	1.229(3)	1.205	0.024
C_1-C_2	1.503(4)	1.518	0.015
C_2-C_3	1.542(2)	1.538	0.004
H_{12}-C_4-H_{13}	106.0(9)	-----	-----
C_2-C_1-C_6	115.3(3)	112.2	3.1
C_1-C_2-C_3	111.5(1)	111.4	0.1
C_3-C_4-C_5	110.8(2)	111.1	0.3
O_7=C_1-C_2-C_3	128.3(50)	-----	-----
C_6-C_1-C_2-C_3	-51.7(50)	-----	-----
C_1-C_2-C_3-C_4	53.0(50)	-----	-----
C_2-C_3-C_4-C_5	-56.3(50)	-----	-----



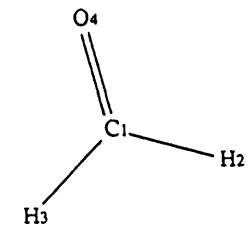
Cycloheptone

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=O_8	1.219(12)	1.207	0.012
C_7-C_1-C_2	117.3(18)	-----	-----
H_{10}-C_6-H_{11}	103.2(26)	-----	-----



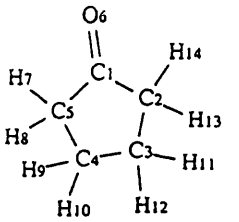
Formaldehyde

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-H_2	1.096(53)	-----	-----
O_4••H_2	2.053(54)	-----	-----
C_1=O_4	1.209(3)	1.196	0.013



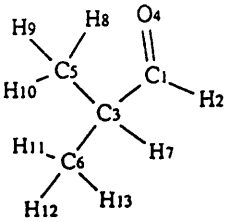
Cyclopentone

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.531(4)	1.500	0.031
C_2-C_3	1.542(3)	1.527	0.015
C_1=O_6	1.213(4)	1.203	0.010
C_2-C_1-C_5	108.6(2)	107.6	1.0



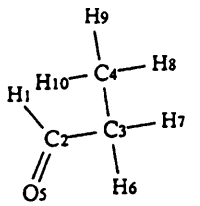
Isopropyl Carboxaldehyde

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=O_4	1.206(12)	-----	-----
C_3-C_5-H_8	110.8(32)	-----	-----
C_3-C_1=O_4	123.3(19)	134.3	11.0



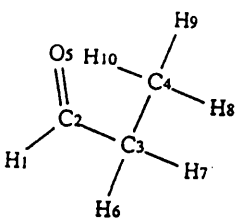
Skew Propanal

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2=O_5	1.212(4)	-----	-----
C_2-C_3	1.512(9)	-----	-----
C_3-C_4	1.569(45)	-----	-----
O_5=C_2-C_3	125.1(3)	126.0	0.9
C_2-C_3-C_4	110.2(4)	112.6	2.4
O_5=C_2-C_3-C_4	123.7(26)	116.6	7.1



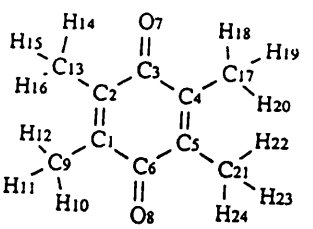
Syn Propanal

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-C_3	1.513(9)	-----	-----
C_2=O_5	1.209(4)	1.199	0.010
O_5=C_2-C_3	124.5(3)	126.5	2.0
C_2-C_3-C_4	113.8(4)	115.5	1.7



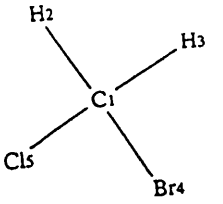
Tetramethyl-p-benzoquinone

Molecular Property	Exptl Value	Calc Value	Outside Error
C_9-H_{10}	1.102(18)	-----	-----
C_1=C_2	1.352(8)	-----	-----
C_3=O_7	1.229(8)	1.202	0.027
C_2-C_3	1.491(11)	1.479	0.012
C_1-C_9	1.504(12)	1.523	0.019
C_2-C_3-C_4	120.8(0.8)	113.9	6.9



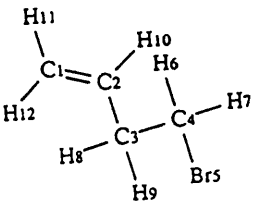
Bromo Chloro Methane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-Br_4	1.928(7)	-----	-----
C_1-Cl_5	1.755(8)	1.767	0.012
H_2-C_1-Br_4	109.3(23)	-----	-----
Cl_5-C_1-Br_4	113.3(6)	110.3	3.0



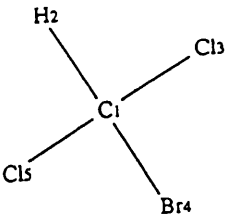
4-Bromo But-1-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2	1.339(10)	-----	-----
C_2-C_3	1.508(7)	-----	-----
C_4-Br_5	1.945(8)	-----	-----
C_3-C_4	1.528(7)	1.580	0.008
C_1=C_2-C_3	122.9(32)	-----	-----
C_2-C_3-C_4	111.5(27)	-----	-----
C_3-C_4-Br_5	111.2(7)	109.7	1.5
C_2-C_3-C_4-Br_5	180.0(50)	-----	-----
C_1=C_2-C_3-C_4	120.0(36)	102.0	18.0



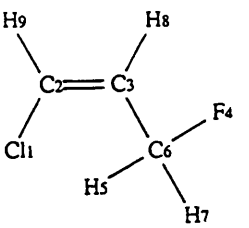
Bromo Dichloro Methane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-Br_4	1.929(9)	-----	-----
C_1-Cl_3	1.758(4)	1.767	0.009
Cl_3-C_1-Cl_5	111.7(7)	108.6	3.1
Cl_3-C_1-Br_4	111.3(3)	109.9	1.4
H_2-C_1-Br_4	107.4(5)	110.9	3.5
H_2-C_1-Cl_3	107.4(5)	108.7	1.3



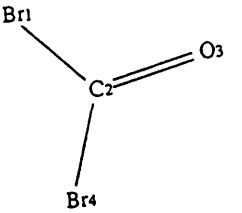
Cis 1-Chloro-3-Fluoro-Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2=C_3	1.347(14)	-----	-----
C_3-C_6	1.493(14)	-----	-----
C_2-Cl_1	1.737(5)	1.747	0.010
C_6-F_4	1.388(9)	1.343	0.045
C_3=C_2-Cl_1	122.7(6)	-----	-----
C_3=C_2-C_6	125.4(8)	123.7	1.7
C_3-C_6-F_4	112.2(8)	109.8	2.4



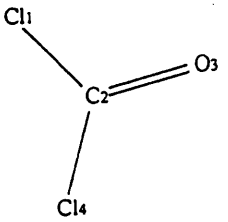
Carbonyl Dibromide

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_2=O_3$	1.178(9)	1.196	0.018
C_2-Br_1	1.923(5)	1.911	0.012
Br_1-C_2-Br_3	112.3(4)	-----	-----



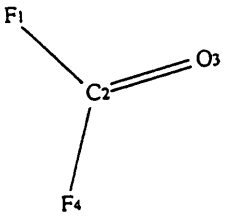
Carbonyl Dichloride

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_2=O_3$	1.177(3)	1.196	0.019
C_2-Cl_1	1.737(2)	1.743	0.006
Cl_1-C_2-Cl_3	111.9(1)	111.7	0.2



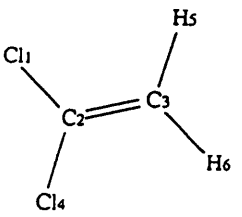
Carbonyl Difluoride

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_2=O_3$	1.172(2)	1.196	0.024
C_2-F_1	1.316(1)	1.327	0.011
F_1-C_2-F_3	107.7(1)	108.6	0.9



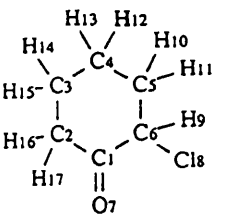
1,1-Dichloro Ethylene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_3-H_5	1.088(11)	-----	-----
$C_2=C_3$	1.329(3)	1.325	0.004
C_2-Cl_1	1.725(2)	1.743	0.016
Cl_1-C_2-Cl_4	114.1(2)	116.7	2.6
H_5-C_3-H_6	121.4(7)	114.9	6.5



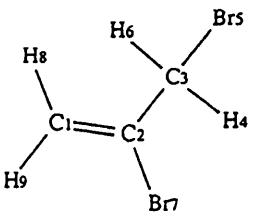
3-Chloro Cyclohexanone

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=O_7$	1.211(2)	1.205	0.006
C_6-Cl_8	1.799(4)	1.784	0.015
$O_7=C_1$-C_2	124.6(5)	124.1	0.005
C_1-C_2-Cl_8	110.2(5)	109.7	0.005
C_3-C_4-C_5-C_6	-55.9(21)	-----	-----
C_2-C_1-C_6-C_5	-48.4(15)	-55.9	7.5
C_1-C_6-C_5-C_4	53.7(7)	55.0	1.3



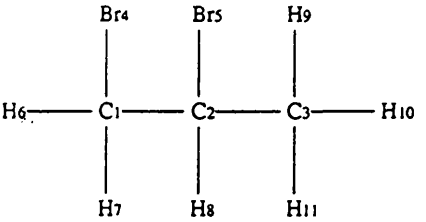
Gauche 2,3-Dibromo Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.333(13)	-----	-----
C_2-C_3	1.480(14)	-----	-----
C_2-Br_7	1.904(17)	1.924	0.020
C_3-Br_5	1.940(19)	2.017	0.077
$C_1=C_2-Br_7$	120.1(44)	-----	-----
$C_1=C_2-C_3$	124.2(17)	-----	-----
$C_2-C_3-Br_5$	113.4(35)	-----	-----



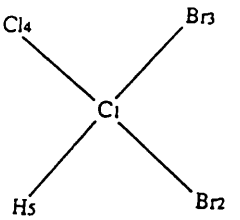
1,2-Dibromo Propane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-Br_4	1.960(7)	-----	-----
$Br_4-C_1-C_2$	110.8(14)	-----	-----
$Br_5-C_2-C_1$	108.6(11)	-----	-----
$Br_5-C_2-C_3$	108.6(11)	-----	-----
$C_1-C_2-C_3$	109.2(24)	113.1	3.9
$Br_4-C_1-C_2-Br_5$	165.0(40)	176.6	11.6



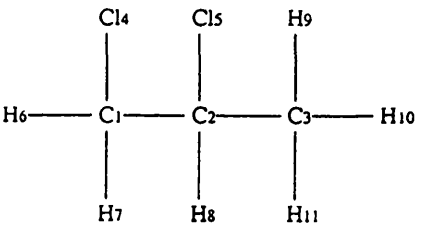
Dibromo Chloro Methane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-Cl_4	1.749(10)	1.767	0.018
C_1-Br_2	1.930(4)	1.923	0.007
$Br_2-C_1-Br_3$	110.0(5)	-----	-----
$H_5-C_1-Cl_4$	107.4(6)	-----	-----
$Cl_4-C_1-Br_2$	111.7(6)	109.2	2.5
$H_5-C_1-Br_2$	107.4(6)	110.1	2.6



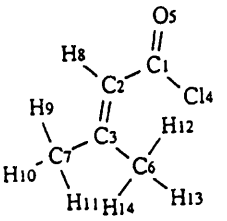
1,2-Dichloro Propane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-Cl_4	1.793(5)	1.773	0.020
$Cl_4-C_1-C_2$	111.2(9)	-----	-----
$Cl_5-C_2-C_1$	108.0(11)	-----	-----
$C_1-C_2-C_3$	114.4(16)	-----	-----
$Cl_5-C_2-C_3$	109.7(11)	108.0	1.7
$Cl_4-C_1-C_2-Cl_5$	165.0(40)	176.6	11.6



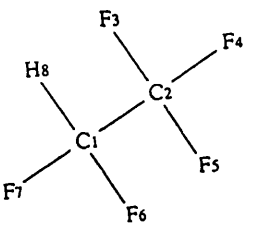
3-Methyl-2-Butenoyl Chloride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.467(8)	-----	-----
C_1=O_5	1.182(4)	1.197	0.015
C_2=C_3	1.333(7)	1.358	0.025
C_3-C_7	1.496(5)	1.521	0.025
C_1-Cl_4	1.800(5)	1.751	0.049
C_2=C_3-C_7	121.3(12)	-----	-----
C_2=C_3-C_6	122.5(10)	-----	-----
C_2=C_2-H_8	120.7(75)	-----	-----
C_2-C_1=O_5	128.3(6)	129.9	1.6
C_2-C_1-Cl_4	111.7(4)	112.2	0.5
C_1-C_2=C_3	127.1(8)	128.6	1.5
C_1-C_2-H_8	114.4(13)	110.8	2.6



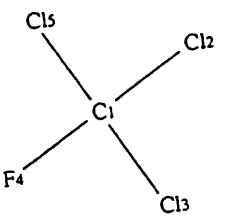
Pentafluoro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.525(4)	1.513	0.012
C_1-F_6	1.347(2)	1.344	0.003
C_2-F_3	1.327(2)	1.344	0.017



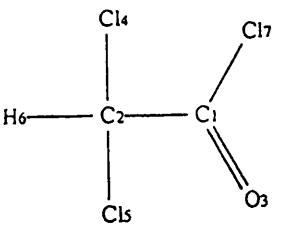
Trichloro Fluoro Methane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-F_4	1.342(3)	-----	-----
C_1-Cl_2	1.761(1)	1.767	0.006
Cl_2-C_1-Cl_3	110.5(1)	110.8	0.3



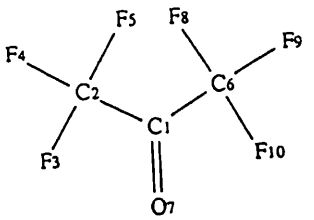
Dichloro Acetylchloride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-H_6	1.062(49)	-----	-----
C_1-Cl_7	1.752(9)	-----	-----
C_2-Cl_4	1.771(4)	-----	-----
C_1=O_3	1.189(3)	1.198	0.009
C_1-C_2	1.535(8)	1.485	0.050
C_2-C_1-Cl_7	113.9(59)	-----	-----
C_1-C_2-Cl_4	109.5(15)	-----	-----
Cl_4-C_2-H_6	108.0(15)	-----	-----
C_2-C_1=O_3	123.3(13)	125.2	1.9
Cl_4-C_2-Cl_5	111.7(5)	108.2	3.5
H_6-C_2-C_1=O_3	138.2(51)	116.9	21.3



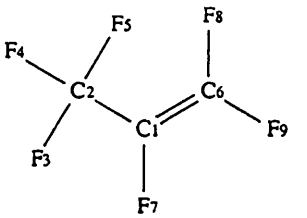
Hexafluoro Acetone

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-F_3	1.335(2)	1.344	0.009
C_1-C_2	1.549(8)	1.501	0.048
C_1=O_7	1.246(14)	1.205	0.041
C_1-C_2-F_3	110.3(3)	112.2	1.9
C_2-C_1-C_6	121.4(4)	112.1	9.3
F_3-C_2-C_1=O_7	36.6(11)	31.1	5.5



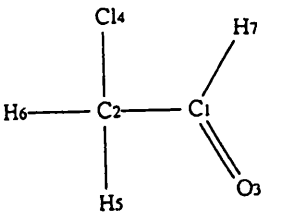
Hexafluoro Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_6	1.329(3)	1.332	0.003
C_1-C_2	1.513(3)	1.492	0.021
F_8-C_6=C_1	123.9(14)	-----	-----
F_7-C_1=C_6	120.0(50)	-----	-----
C_1-C_2-F_3	110.3(15)	-----	-----
C_2-C_1=C_6	127.8(7)	125.6	2.2



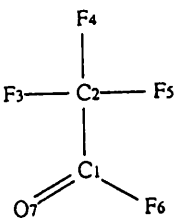
Chloro Acetaldehyde

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-Cl_4	1.782(4)	-----	-----
C_1=O_3	1.206(3)	1.198	0.008
C_1-C_2	1.521(3)	1.502	0.019
C_2-C_1-H_7	112.4(38)	-----	-----
C_1-C_2-H_5	110.3(15)	-----	-----
C_1-C_1=O_3	123.3(6)	124.8	1.5
C_1-C_2-Cl_4	110.4(3)	112.9	2.5
Cl_4-C_2-C_1=O_3	180.0(50)	-----	-----



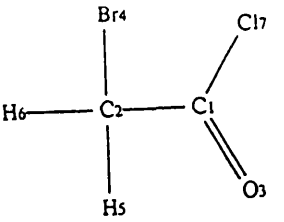
Trifluoro Acetyl Fluoride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.522(6)	1.500	0.022
C_1=O_7	1.158(7)	1.200	0.042
C_2-C_1=O_7	129.0(20)	-----	-----
C_1-C_2-F_3	109.5(5)	112.1	2.6



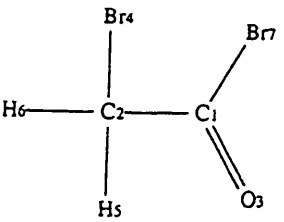
Bromo Acetylchloride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-H_6	1.086(62)	-----	-----
C_1=O_3	1.188(9)	1.200	0.012
C_1-C_2	1.519(18)	1.493	0.026
C_1-Cl_7	1.789(11)	1.744	0.045
C_2-Br_4	1.935(12)	1.959	0.024
C_2-C_1=O_3	127.6(13)	-----	-----
C_1-C_2-Br_4	111.0(15)	-----	-----
C_2-C_1-Cl_7	111.3(11)	111.0	1.3
Br_4-C_2-C_1-Cl_7	110.0(50)	123.2	13.2



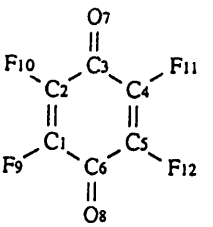
Bromo Acetylbromide

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-H_5	1.110(88)	-----	-----
C_1-C_2	1.513(20)	-----	-----
C_1=O_3	1.175(13)	1.200	0.025
C_1-Br_7	1.987(20)	1.918	0.069
C_2-Br_4	1.915(20)	1.957	0.042
C_2-C_1-Br_7	110.7(15)	-----	-----
C_1-C_2-Br_4	111.7(18)	-----	-----
C_2-C_1=O_3	129.4(17)	125.8	3.6
Br_4-C_2-C_1=O_3	105.0(50)	122.9	17.9



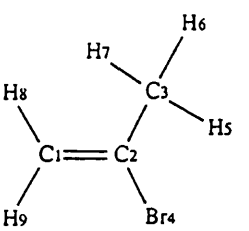
Tetrafluoro-p-benzoquinone

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2	1.339(12)	-----	-----
C_3=O_7	1.211(6)	1.199	0.012
C_2-C_3	1.489(5)	1.468	0.021
C_1-F_9	1.323(5)	1.332	0.009
C_1=C_2-C_3	116.8(7)	123.3	6.5
C_6-C_1-F_9	116.1(7)	118.0	1.7



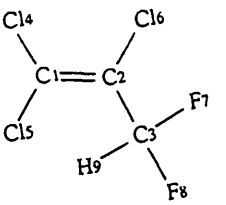
2-Bromo Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2	1.345(4)	1.334	0.011
C_2-C_3	1.499(5)	1.492	0.007
C_2-Br_4	1.906(3)	1.910	0.004
C_1=C_2-C_3	127.2(5)	124.8	2.4
C_1=C_2-Br_4	118.6(6)	120.7	2.1



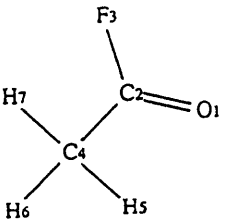
1,1,2-Trichloro 3,3-Difluoro Prop-1-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-C_3	1.509(19)	-----	-----
C_1=C_2	1.365(18)	1.331	0.034
C_3-F_7	1.353(11)	1.341	0.012
C_1=C_2-C_3	124.5(9)	-----	-----
C_2=C_1-Cl_5	123.3(6)	-----	-----
C_1=C_2-Cl_6	120.2(10)	-----	-----
C_2-C_3-F_7	111.0(12)	-----	-----
C_2=C_1-Cl_4	121.8(6)	120.8	1.0



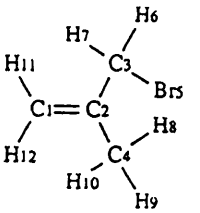
Acetylfluoride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_4-H_5	1.101(4)	-----	-----
C_2=O_1	1.185(2)	1.199	0.014
C_2-F_3	1.362(2)	1.331	0.031
C_2-C_4	1.505(2)	1.500	0.005
C_4-C_2-F_3	110.5(5)	-----	-----
O_1=C_2-F_3	120.7(4)	121.9	1.2
H_5-C_4-H_6	109.3(6)	106.9	2.4



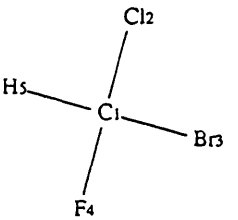
Gauche 3-Bromo 2-Methyl Prop-1-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=C_2	1.331(9)	1.347	0.016
C_2-C_3	1.484(6)	1.514	0.030
C_3-Br_5	1.965(6)	1.955	0.010
C_1=C_2-C_3	121.5(7)	123.0	1.5
C_1=C_2-C_4	120.8(7)	122.4	1.6
C_2-C_3-Br_5	112.2(5)	111.4	0.8
C_1=C_2-C_3-Br_5	112.5(22)	123.5	11.0



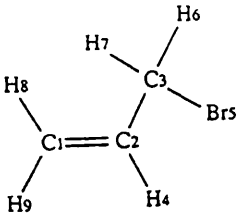
Bromo Chloro Fluoro Methane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-Br_3	1.927(6)	-----	-----
C_1-F_4	1.348(5)	1.341	0.007
C_1-Cl_2	1.753(5)	1.767	0.014
F_4-C_1-Br_3	109.2(6)	-----	-----
Cl_2-C_1-Br_3	111.5(4)	-----	-----
F_4-C_1-Cl_2	109.6(6)	107.5	2.1



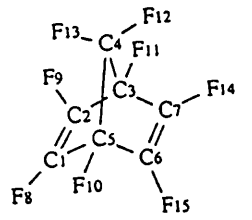
Gauche 3-Bromo Prop-1-ene

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.335(7)	-----	-----
C_2-C_3	1.485(8)	1.498	0.013
C_3-Br_5	1.961(6)	1.936	0.025
$C_2-C_3-Br_5$	111.5(8)	110.4	1.1
$C_1=C_2-C_3-Br_5$	117.0(50)	142.0	25.0



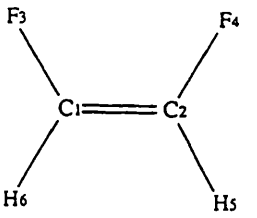
Perfluoro Norbornadiene

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.320(20)	-----	-----
C_2-C_3	1.520(10)	1.496	0.024
C_3-C_4	1.560(20)	1.519	0.041
C_1-F_8	1.310(10)	1.329	0.019
$C_1=C_2-C_3$	107.0(10)	-----	-----
$C_2-C_3-C_7$	108.0(10)	-----	-----
$C_3-C_4-C_5$	90.0(10)	93.2	3.2
$C_1=C_2-F_9$	130.0(10)	128.5	1.5
$F_{12}-C_4-F_{13}$	108.0(10)	105.5	2.5
$C_2-C_3-C_4$	99.0(10)	97.9	1.1



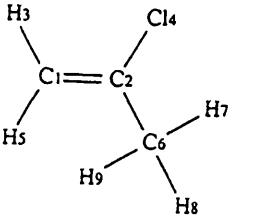
Cis 1,2-Difluoro Ethene

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.330(11)	-----	-----
C_1-F_3	1.342(5)	1.328	0.014
C_1-H_6	1.103(6)	1.089	0.014
$C_1=C_2-F_4$	122.0(2)	121.7	0.3
$C_1=C_2-H_5$	124.1(4)	124.7	0.6



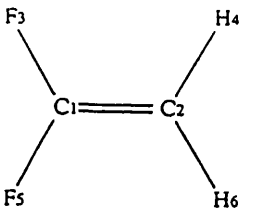
2-Chloro Propene

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.336(3)	-----	-----
C_2-C_3	1.491(3)	-----	-----
C_2-Cl_4	1.743(2)	-----	-----
C_1-H_1	1.088(6)	-----	-----
$C_1=C_2-C_3$	126.7(4)	125.4	1.3
$C_1-C_2-Cl_4$	119.1(4)	120.4	1.3



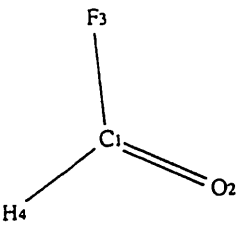
1,1-Difluoro Ethene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-H_4	1.091(10)	-----	-----
C_1=C_2	1.340(6)	1.330	0.010
C_1-F_3	1.315(3)	1.329	0.014
C_2=C_1-F_3	124.7(3)	123.3	1.4
C_1=C_2-H_4	119.0(4)	122.8	3.8



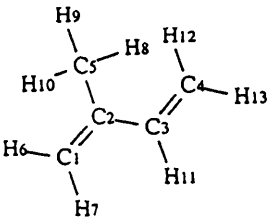
Formylfluoride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=O_2	1.188(4)	1.196	0.008
C_1-F_3	1.346(3)	1.327	0.019
C_1-H_4	1.110(20)	1.086	0.024
O_2=C_1-H_4	130.0(40)	-----	-----
O_2=C_1-F_3	122.3(2)	124.0	1.7



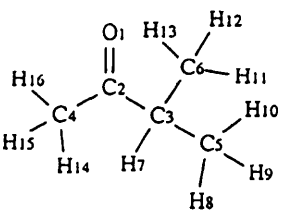
Isoprene

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2-C_5	1.512(3)	-----	-----
C_2-C_3	1.463(3)	1.479	0.016
C_1=C_2	1.340(3)	1.345	0.005
C_1=C_2-C_3	121.4(3)	-----	-----
H_7-C_1=C_2	124.3(10)	123.1	1.2
C_2-C_3=C_4	127.3(3)	125.3	2.0
C_1=C_2-C_5	121.0(2)	120.3	0.7
H_{11}-C_3=C_4	123.4(12)	117.5	5.9
C_2-C_5-H_8	109.1(14)	113.5	4.4



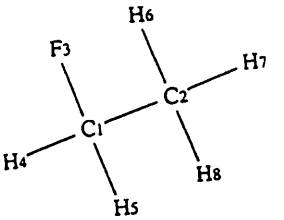
Skew 3-Methyl Butan-2-one

Molecular Property	Exptl Value	Calc Value	Outside Error
C_2=O_1	1.217(2)	1.204	0.013
C_2-C_3-C_6	113.4(12)	-----	-----
C_5-C_3-C_6	111.7(21)	-----	-----
C_3-C_2-C_4	118.0(2)	115.0	3.0
O_1=C_2-C_3	120.8(14)	123.0	2.2
C_2-C_3-C_5	108.0(10)	110.2	2.3
O_1=C_2-C_3-H_7	140.0(40)	158.4	16.4



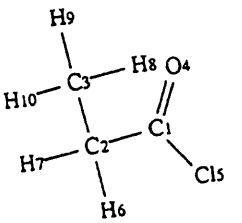
Monofluoro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.502(5)	1.525	0.023
C_1-F_3	1.397(4)	1.344	0.053
C_2-C_1-F_3	110.4(2)	109.8	0.6



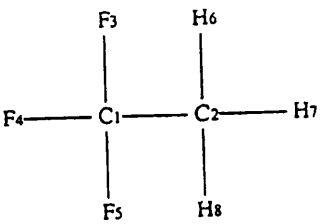
Propionyl Chloride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1=O_4	1.187(5)	1.199	0.012
C_1-Cl_5	1.795(5)	1.749	0.046
C_2-C_1=O_4	127.0(7)	-----	-----
H_6-C_2-H_7	105.4(98)	-----	-----
C_2-C_1-Cl_5	112.1(4)	110.2	1.9
C_1-C_2-C_3	112.7(7)	116.4	3.7
C_1-C_2-H_6	102.0(41)	107.7	5.7



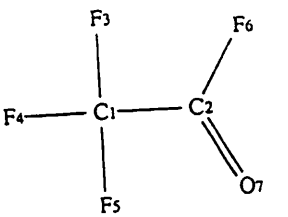
1,1,1-Trifluoro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.494(3)	1.513	0.019
C_1-F_3	1.340(2)	1.344	0.004
C_2-H_6	1.081(7)	1.098	0.017
C_2-C_1-F_3	111.9(2)	-----	-----
C_1-C_2-H_6	112.0(10)	111.0	1.0



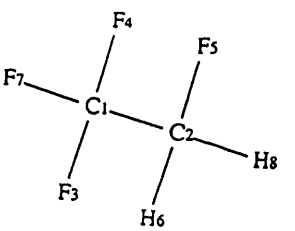
Trifluoro Acetylfluoride

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.522(6)	1.500	0.022
C_1-F_3	1.324(2)	1.345	0.021
C_2=O_7	1.158(7)	1.200	0.042
C_1-C_2=O_7	129.0(20)	-----	-----



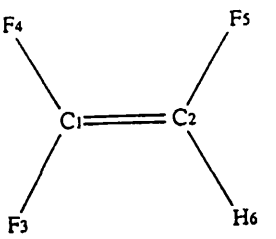
1,1,1,2-Tertafluoro Ethane

Molecular Property	Exptl Value	Calc Value	Outside Error
C_1-C_2	1.501(4)	1.513	0.012
C_2-H_6	1.077(15)	1.098	0.021
C_2-F_5	1.389(6)	1.345	0.044
C_1-C_2-H_6	106.1(12)	111.7	5.6
C_1-C_2-F_5	112.3(4)	110.6	1.7
F_3-C_1-F_4	108.6(2)	106.8	1.8
F_5-C_2-H_6	111.4(15)	106.8	4.6



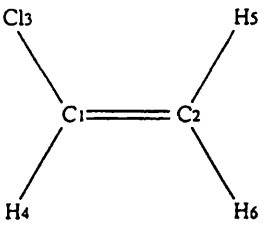
Trifluoro Ethene

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.338(12)	-----	-----
C_1-F_4	1.315(18)	-----	-----
C_2-F_5	1.314(22)	-----	-----
C_1-F_3	1.309(18)	1.329	0.020
C_2-H_6	1.074(10)	1.087	0.013
$C_1=C_2$-H_6	120.0(17)	-----	-----
$C_2=C_1$-F_3	124.4(24)	-----	-----
$C_2=C_1$-F_4	122.6(24)	-----	-----
$C_1=C_2$-F_5	120.0(7)	121.9	1.9



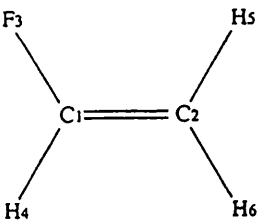
Vinyl Chloride

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.342(4)	1.331	0.011
C_1-Cl_3	1.730(4)	1.742	0.012
$C_2=C_1$-Cl_3	122.5(2)	121.4	1.1
$C_2=C_1$-H_4	123.8(15)	122.2	1.6
$C_1=C_2$-H_5	121.1(7)	123.7	2.6



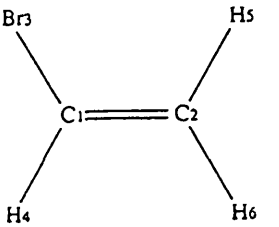
Vinyl Fluoride

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.330(18)	-----	-----
C_1-F_3	1.351(15)	1.329	0.022
$C_2=C_1$-F_3	121.5(2)	-----	-----
$C_2=C_1$-H_4	131.0(30)	124.8	6.2
$C_1=C_2$-H_5	120.0(20)	123.2	3.2



Vinyl Bromide

Molecular Property	Exptl Value	Calc Value	Outside Error
$C_1=C_2$	1.348(8)	1.331	0.017
C_1-Br_3	1.881(7)	1.907	0.026
$C_2=C_1$-Br_3	122.8(3)	121.6	1.2
$C_2=C_1$-H_4	123.1(10)	121.2	1.9
$C_1=C_2$-H_5	120.4(7)	121.8	1.4



7.5.1.2 Structure Residual Factors

Figures 7.47 to 7.49 show the change in the various structural residual factors between the beginning and the end of the final force field optimisation. The following table 7.7 is a summary of the standard deviations for the initial and final force fields for this optimisation.

Table 7.7 Change in the Standard Deviations for Structural Properties

Type of Property	Initial	Final
Distance	3.537	3.450
Valence Angle	3.584	3.119
Torsion Angle	1.848	2.002

It can be seen that the effect of the optimisation has been to narrow the distribution of, mainly, the valence angle at the expense of the torsion angle, but that the movement has been very small for all the property types. Such small movements are to be expected as the parameters used at the beginning were the result of earlier optimisations on similar force field configurations.

Figure 7.47 Distance Residual Factors for Final Optimisation

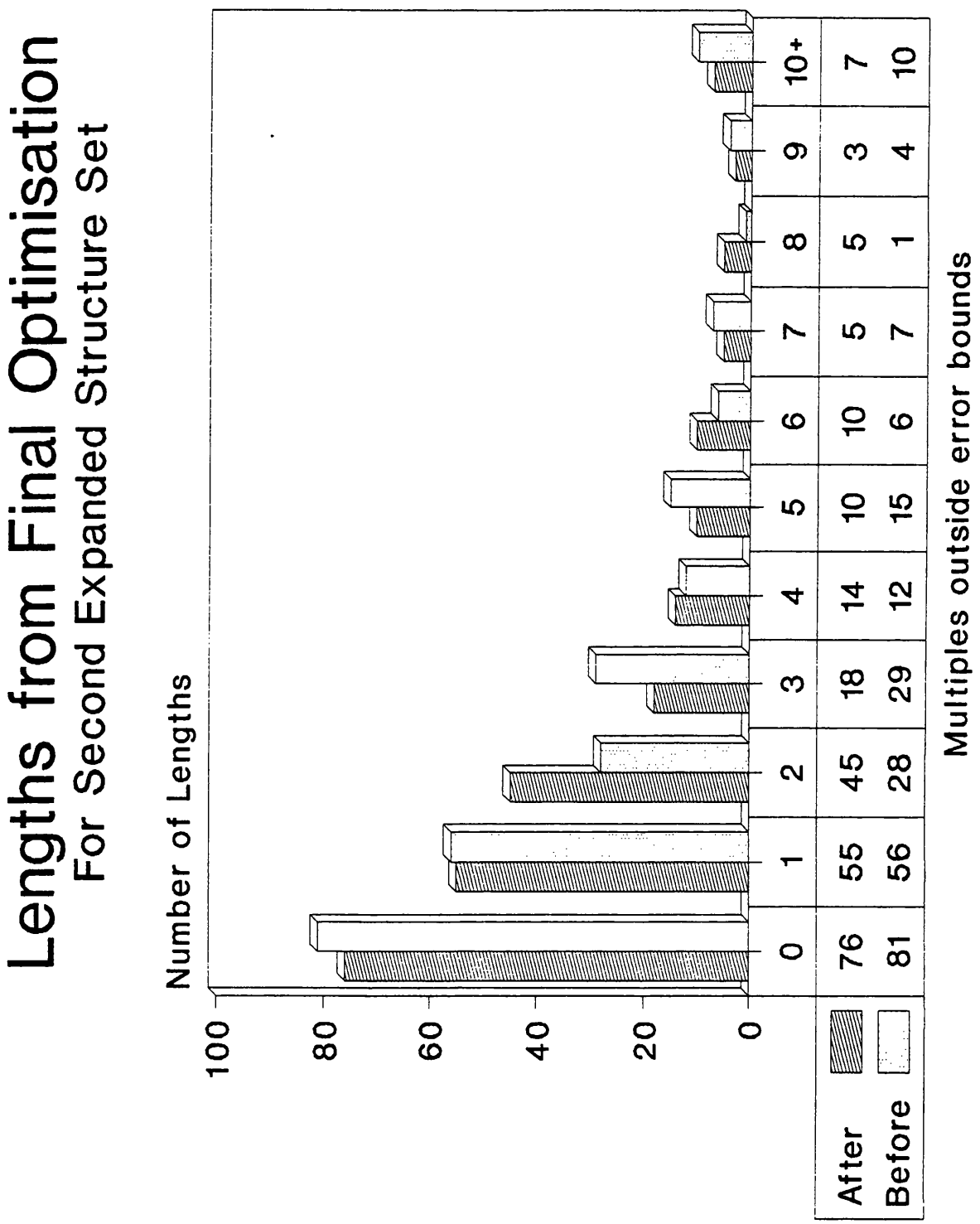


Figure 7.48 Valence Angle Residual Factors for Final Optimisation

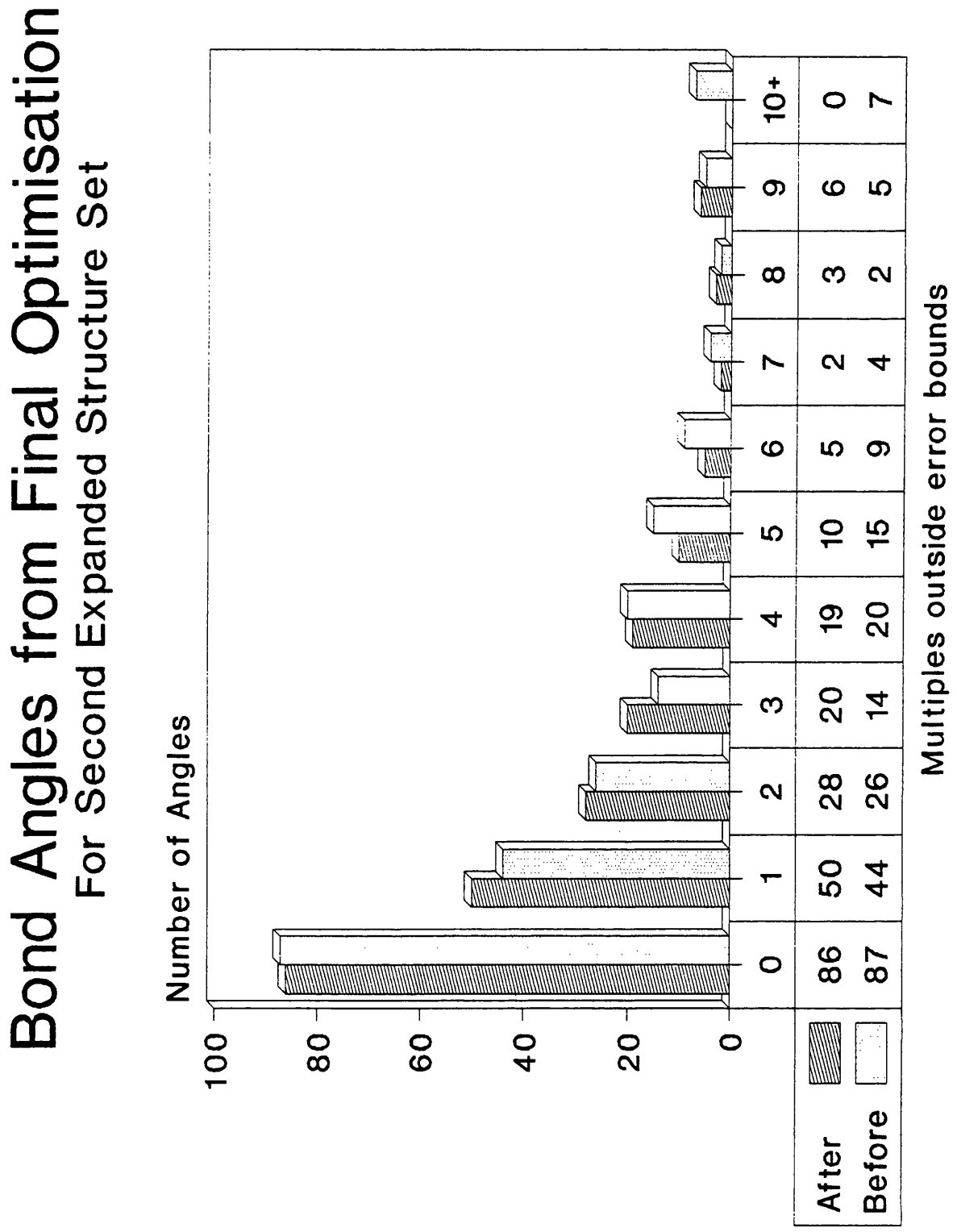
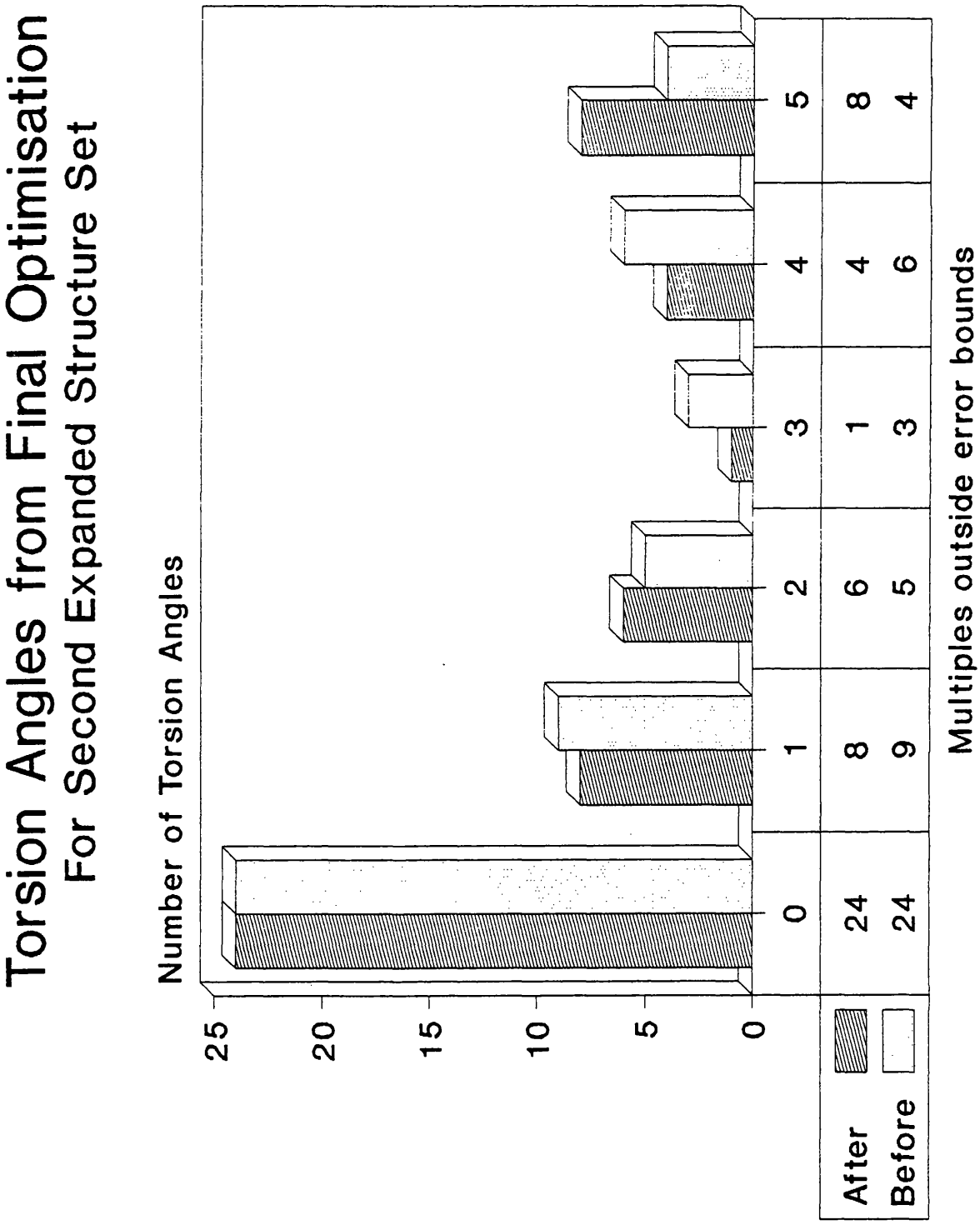


Figure 7.49 Torsion Angle Residual Factors for Final Optimisation



7.5.1.3 Energy Difference Results

There was a total of 16 energy differences used in the final optimisation. The results, taken from the produced COMPARE file, are as follows. Note that all the numbers were outputted in the same format so trailing zero's should be discounted when considering the number of significant figures for the experimental data:

```
BETWEEN GAUCHE-BUTANE
  AND  ANTI-BUTANE

CALC    0.826 EXPT    0.824 ERROR    0.143 DIF    0.000

BETWEEN CIS BUT 2 ENE
  AND  TRANS BUT 2 ENE

CALC    1.304 EXPT    1.130 ERROR    0.380 DIF    0.000

BETWEEN BOAT-CHAIR BICYCLO[3.3.1] NONANE
  AND  TWIN-CHAIR BICYCLO[3.3.1] NONANE

CALC    2.911 EXPT    3.050 ERROR    0.150 DIF    0.000

BETWEEN GAUCHE-1,2 DIFLUORO ETHANE
  AND  ANTI 1,2-DIFLUORO ETHANE

CALC    1.019 EXPT    0.770 ERROR    0.400 DIF    0.000

BETWEEN GAUCHE - 1,1,2,2 TETRA FLUORO ETHANE
  AND  ANTI - 1,1,2,2 TETRA FLUORO ETHANE

CALC    1.020 EXPT    1.170 ERROR    0.200 DIF    0.000

BETWEEN GAUCHE 1 BROMO PROPANE
  AND  ANTI 1 BROMO PROPANE

CALC    0.423 EXPT    0.110 ERROR    0.030 DIF    0.283

BETWEEN GAUCHE 1 CHLORO 2 BROMO ETHANE
  AND  ANTI 1 CHLORO 2 -BROMO ETHANE

CALC    1.346 EXPT    1.400 ERROR    0.200 DIF    0.000

BETWEEN SYN TRANS-2-BUTENOYL CHLORIDE
  AND  ANTI TRANS-2-BUTENOYL CHLORIDE

CALC   -0.089 EXPT    0.100 ERROR    0.200 DIF    0.000

BETWEEN ANTI-SYN FUMARYL CHLORIDE
  AND  ANTI-ANTI FUMARYL CHLORIDE

CALC    0.044 EXPT    0.600 ERROR    0.200 DIF    0.356

BETWEEN SYN-SYN FUMARYL CHLORIDE
  AND  ANTI-ANTI FUMARYL CHLORIDE

CALC    0.044 EXPT    0.700 ERROR    0.200 DIF    0.456
```

BETWEEN CHLORO ACETYL CHLORIDE (GAUCHE)
AND CHLORO ACETYL CHLORIDE (ANTI)

CALC 0.204 EXPT 1.300 ERROR 0.400 DIF 0.696

BETWEEN GAUCHE 2-BROMO-3-CHLORO-PROPENE
AND ANTI 2-BROMO-3-CHLORO PROPENE

CALC 1.241 EXPT 3.000 ERROR 1.000 DIF 0.759

BETWEEN GAUCHE 1,2-DIBROMO ETHANE
AND ANTI 1,2-DIBROMO ETHANE

CALC 1.413 EXPT 2.200 ERROR 0.140 DIF 0.647

BETWEEN GAUCHE 1,2-DICHLORO ETHANE
AND ANTI 1,2-DICHLORO ETHANE

CALC 1.305 EXPT 1.050 ERROR 0.100 DIF 0.155

BETWEEN GAUCHE 2,3-DICHLORO-1-PROPENE
AND ANTI 2,3-DICHLORO-1-PROPENE

CALC 1.262 EXPT 0.700 ERROR 0.300 DIF 0.262

BETWEEN GAUCHE 2,3-DICHLORO-1,3-BUTADIENE
AND ANTI 2,3-DICHLORO-1,3-BUTADIENE

CALC 1.712 EXPT 2.200 ERROR 1.100 DIF 0.000

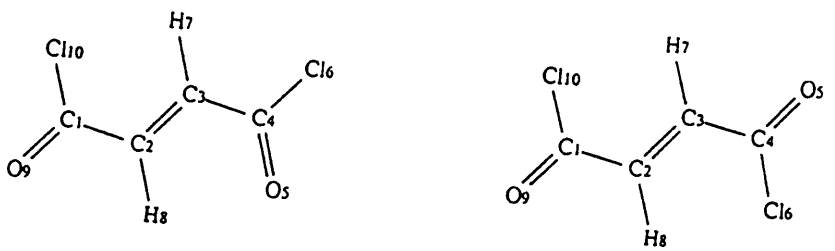
BETWEEN GAUCHE 1,1,2,2-TETRAFLUORO ETHANE
AND ANTI 1,1,2,2-TETRAFLUORO ETHANE

CALC 1.042 EXPT 1.170 ERROR 0.100 DIF 0.028

One obvious problem that can be seen from these results is that the energy difference is the same for both the Anti-Syn/Anti-Anti and the Syn-Syn/Anti-Anti Fumaryl Chloride cases and that this calculated value is considerably lower than the experimental value in both of these cases.

Figure 7.50 shows both Anti-Syn and Anti-Anti Fumaryl Chloride.

Figure 7.50 Anti-Syn and Anti-Anti Fumaryl Chloride



Looking at the given structural properties for the Anti-Syn case it be seen that the experimental value of the torsion angle Cl_6-C_4-C_3=C_2 is 155.3(5.5)° while the calculated value is a planar 180.0°. This implies that in real life there is some force causing the twisting of the COCl group that is either not represented or present but not of sufficient magnitude to overcome the two fold torsional barrier across the conjugated bond. This force must either be a torsional term that is not given or a non-bonded repulsion not being accurately modelled. If it is assumed that this is the case then it can be assumed that the energy of this other force should be equal to the rise in energy from the torsional interaction. This, in combination with the final force field parameter for the conjugated bond gives rise to additional energy value of 0.569 Kcal Mol^{-1} giving a total difference almost exactly equal to the experimental value.

7.5.1.4 Energy Difference Residual Factors

Figure 7.51 shows the distribution of the energy difference residual factors before and after the final optimisation. The standard deviations for the results are given in table 7.8.

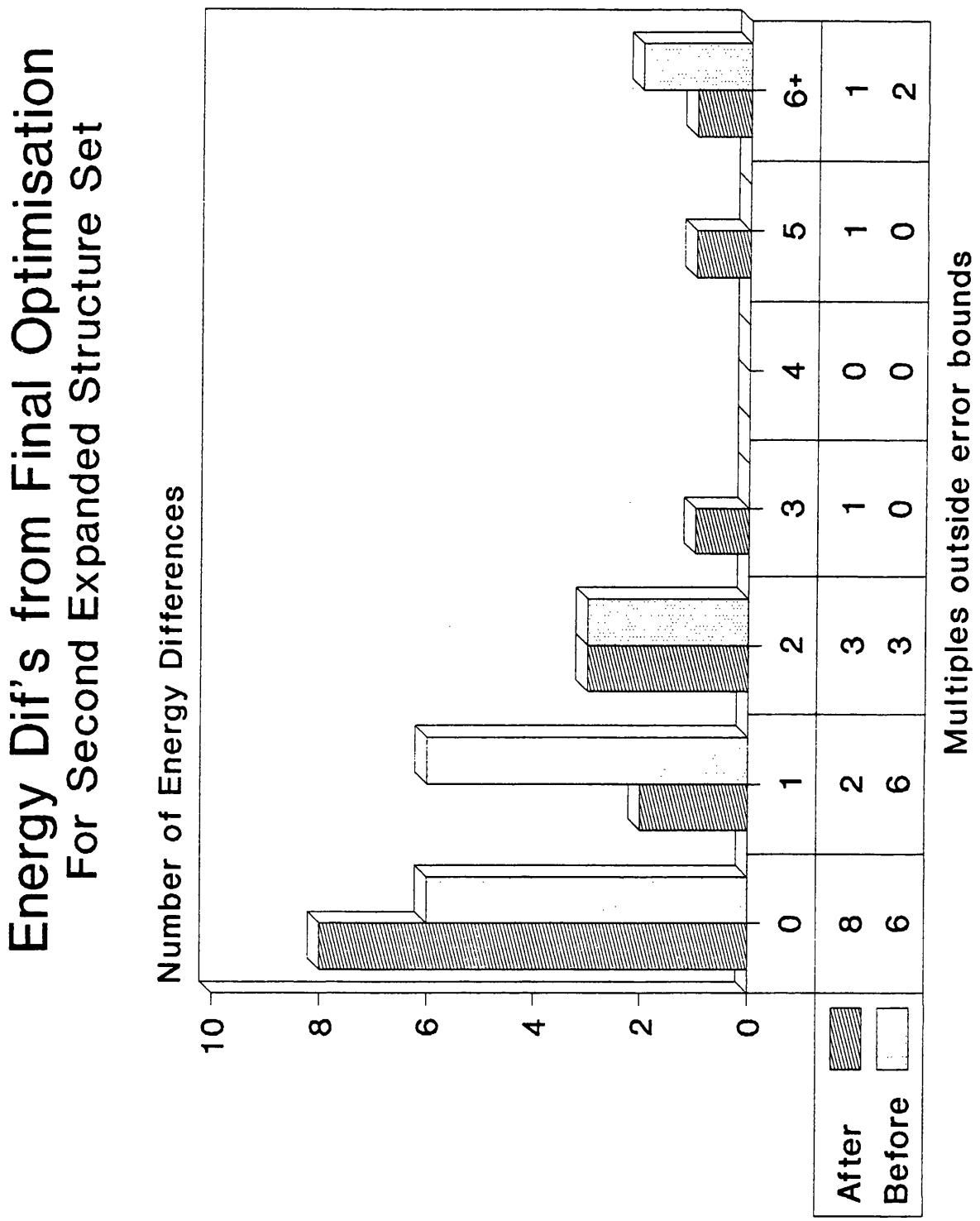
Table 7.8 Change in Standard Deviation for Isomeric Energy Differences

Type of Property	Initial	Final
Energy Difference	4.453	2.717

This shows that of the properties examined so far the greatest improvement has been in the energy difference. This is probably because the energy difference is the highest weighted residual factor. Most of this improvement comes from the energy difference between Gauche and Anti Bromo-Propane, where, because of the associated low error, the residual factor went from 16.9 to 9.43. In absolute terms the movement was slight with the difference between the average absolute difference before and after being only 0.1362 KCal Mol^{-1}.

The overall impression that can be obtained from the energy difference residual factors is that there are cases, such as those given at the end of the previous section, which are giving results that imply that they are failing to represent the situations for which they are being compared to the experimental values. This, taken with the fact that the number of energy differences is relatively small compared to both the total number of parameters being optimised and the number of residual factors for the other optimisation residual factors, means that the choice of such a large weighting factor for this residual factor may have played a significant part in the

Figure 7.51 Isomeric Energy Residual Factors for Final Optimisation



production of chemically unreasonable values to several of the parameter types. The problem with the weighting can be traced back to the comparison of the sizes of the various residual factors. The following is the section of the final COMPARE file given above which show the size of the separate residual factors:

STRUCTURE	:	10.36323
ENERGY	:	0.11694
HEAT OF FORMATION	:	14.11788

In combination with the weightings these result in the following contributions to the total residual factor:

STRUCTURE	:	51.816
ENERGY	:	2.339
HEAT OF FORMATION	:	14.118
TOTAL	:	68.27274

This shows how small the energy difference residual factor is compared to the other residual factors. On the basis of this it could be said that rather than being over-weighted the energy difference residual factor is under-weighted. This appearance is, to an extent, deceptive as it relies on the size of the residual factors and not their ranges. For example, with the data set under investigation and for reasons given in section 7.3.1.1, the distance component of the structural contribution will always tend to make a large contribution to the structure residual factor irrespective of the force field under consideration. A similar problem occurs in the heat of formation residual factor where the empirical nature of the geometric enthalpy increments means that it is unlikely that the overall residual factor for the final structure set will be zero.

So for the other two residual factors there will tend to be some, unspecifiable, lower limit that means that it is difficult to use the values of the residual factors in the assignment of weights. This point, not considered when deciding the weighting factors, may have resulted in the overweighting of the energy difference residual factors.

7.5.1.5 Heats of Formation Results

The following is a list of the Heats of Formation used in the optimisation. For each the following is given:

- The Calculated Value
- The Experimental Value
- The Error for Experimental Value
- The amount by which the calculated value exceeds the error limits of the experimental value.

Note that as with the isomeric energy differences the output format disguises

the number of significant figures in the experimental data. To determine the number of significant figures trailing zero's should be discounted.

Structure	Calc	Expt	Error	Outside Error
CIS BUT 2 ENE	-0.2932	-1.6970	0.2400	1.1638
TRANS BUT 2 ENE	-1.5972	-2.7250	0.2400	0.8878
ADAMANTANE	-34.3296	-32.1700	0.5500	1.6096
CH-CH-BICYCLO(3.3.2.)DECANE	-18.2687	-25.3580	1.7690	5.3203
CIS-HEX-2-ENE	-12.4814	-12.5100	0.3300	0.0000
CYCLOHEXA-1,4-DIENE	28.1499	26.3000	0.1000	1.7499
CYCLOHEXENE	-1.6213	-1.1950	0.1670	0.2593
CYCLOHEPTENE	-3.4072	-2.1990	0.2630	0.9452
CYCLOPENTENE	8.6351	8.1020	0.3350	0.1981
2,3-DIMYTHYL-BUT-1-ENE	-14.4396	-14.9620	0.3100	0.2124
2,3-DIMETHYLBUT-2-ENE	-14.6485	-16.3000	0.2630	1.3885
4,4-DIMETHYL-PENT-1-ENE	-20.4337	-19.5030	0.4780	0.4527
ETHYLENE	14.8785	12.5480	0.0960	2.2345
HEXA-1,5-DIENE	22.5182	20.1000	0.2150	2.2032
ISOBUTENE	-3.2898	-4.0390	0.2150	0.5342
1-METHYL-ADAMANTANE	-42.2161	-41.0130	0.6690	0.5341
1-METHYL-CYCLOHEXENE	-10.7058	-10.3490	0.1910	0.1658
1-METHYL-CYCLOPENTENE	-1.7375	-0.9080	0.1670	0.6625
1-METHYL-DIAMANTANE	-37.6500	-39.8420	0.7890	1.4030
2-METHYL-BUT-2-ENE	-9.1610	-9.9900	0.2630	0.5660
2-METHYL-PENT-2-ENE	-15.4869	-15.9900	0.3590	0.1441
3-METHYL-BUT-1-ENE	-5.7681	-6.5970	0.1910	0.6379
4-METHYL-CIS-PENT-2-ENE	-13.5350	-13.7430	0.2870	0.0000
4-METHYL-TRANS-PENT-2-ENE	-14.0103	-14.6990	0.3350	0.3537
2-METHYL-BUT-1-ENE	-8.7989	-8.4370	0.2390	0.1229
NORBORNADIENE	63.8453	58.7720	0.6930	4.3803
NORBORNANE	-15.5996	-13.1210	1.1230	1.3556
PENT-1-ENE	-5.3907	-5.0910	0.2390	0.0607
PROPENE	6.3608	4.7800	0.1910	1.3898
PENTA-1,4,DIENE	28.3918	25.2390	0.2870	2.8658
PERITRIHYDROQUINACENE	-25.8175	-24.4500	0.8370	0.5305
5-SKEW-CIS-PENT-2-ENE	-6.6204	-6.8600	0.2000	0.0396
5-SKEW-TRANS-PENT-1-ENE	-7.7553	-7.7600	0.2000	0.0000
TRANS,TRANS,TRANS,CYCLODODECA-1,5,9-TRIENE	18.8402	24.2110	1.0280	4.3428
CYCLOHEXANE	-33.0801	-29.5000	0.2000	3.3801
TRANS-HEX-2-ENE	-13.5492	-12.8800	0.3900	0.2792
1,3,5,7-TETRAMETHYL-ADAMANTANE	-66.2211	-67.7340	1.6490	0.0000
2,4,4-TRIMETHYL-PENT-1-ENE	-28.8478	-26.4340	0.3580	2.0558
ACETALDEHYDE	-45.3338	-39.6700	0.1200	5.5438
ACETONE	-55.2147	-51.9000	0.1700	3.1447
2,3 BUTANEDIONE	-78.1300	-78.1300	0.3100	0.0000
CYCLO HEPTONE	-56.6836	-55.1100	0.4300	1.1436
CYCLO HEXANONE	-57.9101	-54.0000	0.5000	3.4101
CYCLOPENTANONE	-49.7386	-45.8800	0.4300	3.4286
FORMALDEHYDE	-37.4695	-25.9400	0.1200	11.4095
1 BROMO-2 CHLORO 1,1,2 TRIFLUORO ETHANE	-166.1726	-167.4000	4.9700	0.0000
CARBONYL BROMIDE	-27.4761	-27.1100	0.1400	0.2261
CARBONIC DICHLORIDE	-51.0535	-52.3300	0.1000	1.1765
CARBONIC DIFLUORIDE	-134.2357	-152.8100	0.3600	18.2143

Structure	Calc	Expt	Error	Outside Error
CARBON TETRAFLUORIDE	-217.8737	-223.0000	0.3300	4.7963
1,1 CHLORO ETHYLENE	-0.7488	0.6200	0.3100	1.0588
CHLOROTRIFLUORO METHANE	-168.3956	-168.7000	0.8800	0.0000
CARBON TETRACHLORIDE	-19.9627	-22.9000	0.2600	2.6773
1,2 DIBROMO-1,CHLORO-1,2,2TRIFLOURO ETHANE	-158.7076	-156.8300	4.9700	0.0000
1,2 DIBROMO PROPANE	-16.8462	-17.0800	0.2600	0.0000
1,2 DICHLORO PROPANE	-33.0342	-38.8800	0.2900	5.5558
DICHLORODIFLUORO METHANE	-118.9180	-114.0200	1.7000	3.1980
TRICHLORO FLUORO METHANE	-69.4404	-64.1000	2.0000	3.3404
ACETYL FLUORIDE	-95.5134	-105.5900	0.7900	9.2866
3-BROMO PROPENE	14.0049	10.8000	1.0000	2.2049
2-CHLORO PROPENE	-3.5872	-5.0200	2.2500	0.0000
1,1 DIFLUORO ETHENE	-83.6397	-80.0100	1.0700	2.5597
ISOPRENE	18.0300	18.0300	0.2600	0.0000
SKEW 3-METHYL-2-BUTANONE	-62.4110	-62.7000	0.2100	0.0790
MONOBROMO ETHANE	-15.9248	-14.7800	0.4100	0.7348
MONOCHLORO ETHANE	-24.0479	-26.7700	0.2600	2.4621
1,1,1-TRIFLUORO ETHANE	-175.9638	-174.5200	4.9900	0.0000
TRIFLUORO ETHENE	-130.2466	-117.1300	0.2400	12.8766
VINYL CHLORIDE	7.6618	8.9100	0.3100	0.9382
VINYL FLUORIDE	-33.7660	-33.1500	0.4100	0.2060
VINYL BROMIDE	19.6521	18.9200	0.4500	0.2821

There are no values that appear to be as bad as the cases highlighted in the Energy Difference section. This is not surprising as the least square's operation required to calculate the geometric enthalpy increments will tend to have a levelling out effect, giving rise to higher errors on those cases where the steric energy is close, while lowering the error on those cases where the correlation is not so good.

7.5.1.6 Heat of Formation Residual Factors

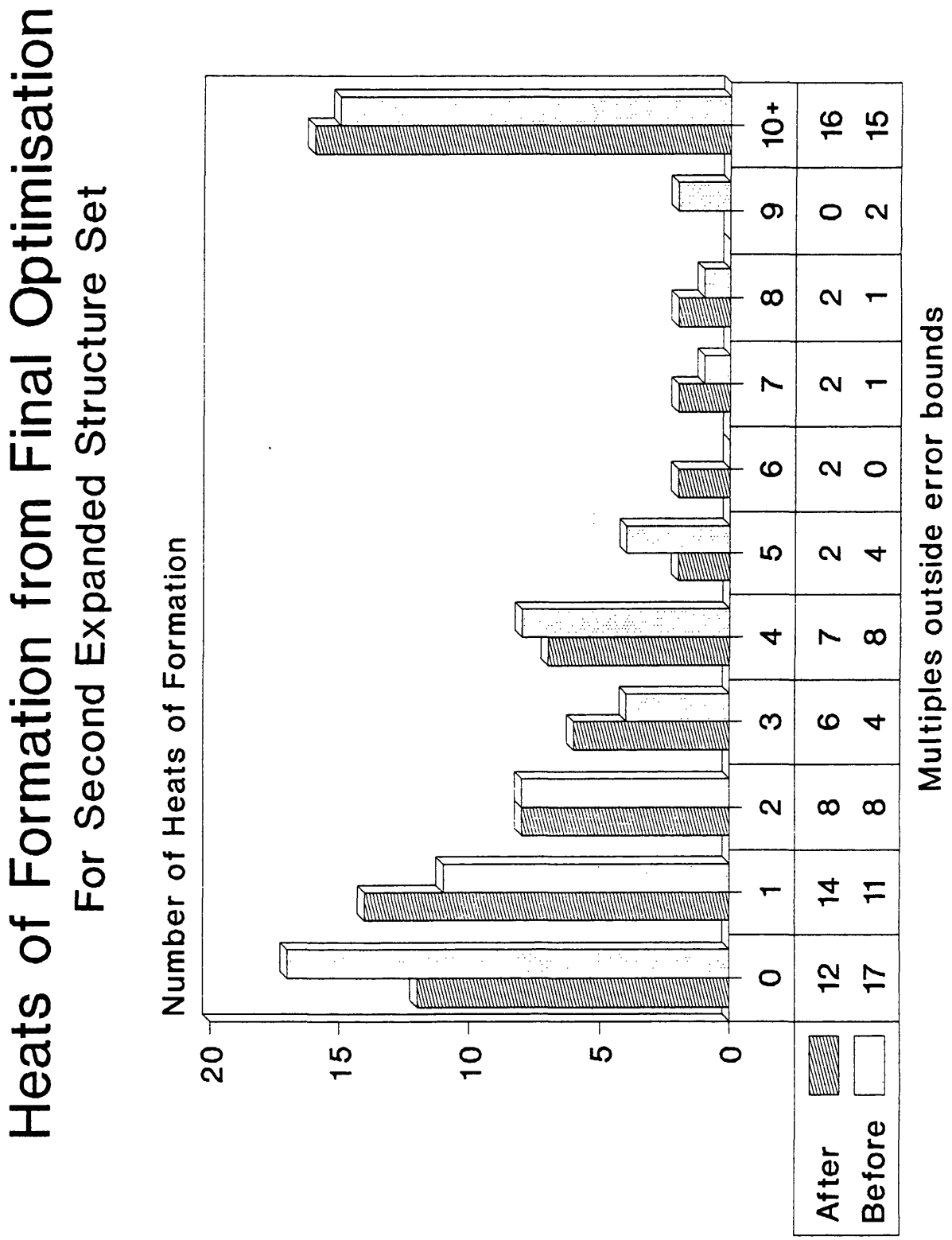
Figure 7.52 shows the distribution of the heat of formation residual factors before and after the final optimisation. The standard deviations for the results are given in table 7.9.

Table 7.9 Change in the Standard Deviation for the Heat of Formation Factors

Type of Property	Initial	Final
Heat of Formation	16.817	16.678

This shows only a slight decrease in the residual factor. As explained above this is to be expected as the use of least squares is likely to minimise the effect of variations in the steric energies on the residual factor. Examining the values it can be seen that the majority of this residual factor comes from the following structures.

Figure 7.52 Heat of Formation Residual Factors for Final Optimisation



Structure	Calc	Expt	Error Outside Error	Residual f'tor
CYCLOHEXANE	-33.0801	-29.5000	0.2000 3.3801	16.9
CYCLOHEXA-1,4-DIENE	28.1499	26.3000	0.1000 1.7499	17.5
ACETONE	-55.2147	-51.9000	0.1700 3.1447	18.5
1,2 DICHLORO PROPANE	-33.0342	-38.8800	0.2900 5.5558	19.2
ETHYLENE	14.8785	12.5480	0.0960 2.2345	23.3
ACETALDEHYDE	-45.3338	-39.6700	0.1200 5.5438	46.2
CARBONIC DIFLUORIDE	-134.2357	-152.8100	0.3600 18.2143	50.6
TRIFLUORO ETHENE	-130.2466	-117.1300	0.2400 12.8766	53.7
FORMALDEHYDE	-37.4695	-25.9400	0.1200 11.4095	95.1

For the first 6 structures the high residual factor is not due to a particularly large difference but to low experimental errors, averaging at 0.162 compared to an average for all the structures of 0.6726. The remaining structures also have below average errors but in these cases the energy difference is more significant. One thing to notice about these are that they all use very few geometric enthalpy increments. Carbonic Difluoride and Formaldehyde use only two and Trifluoro Ethene three. This tendency of the bad results to coincide with structures with few geometric increments means that the conclusion must be that the high residual factors are a result of problems coming from the use of empirical geometric enthalpy increments.

7.5.1.7 Comparison with Other Force Field Configurations

The previous sections give the results of the final optimisation but it is also important to compare these results with those obtained from earlier optimisations that used different force field configurations. This is important as it is necessary to see what happens when mainly reduced parameter options are used together.

In this investigation the seven runs utilised in the initial investigation of the second expanded structure set, with the same weighting, are used. For each of these a note was taken of the final total residual factor, the standard deviations for each property type and the number of parameters that were optimised.

This data is then used to construct the following scatter diagrams that plot the residual factor or standard deviation against the number of parameters in the optimisation for these seven runs and the final optimisation. The final optimisation used a total of 50 optimisable parameters and in the following graphs it is represented by the black square on the far left hand side of the graph. All the other results are indicated using squares containing crosses.

Table 7.10 shows the parameter options which are different for the studies.

All the optimisations for which data is shown had 'full' bond lengths, 'calculated' bond stretching force constants, 'central 2' torsion barriers and out of plane bending parameters.

Table 7.10 Parameter Numbers for Studies Used in Comparison

Angle Value	Angle Bend	Non-Bond	No. of Param's
Full	Full	A6-Exp	170
Full	Full	Cosmic	128
Calculated	Full	Cosmic	94
Full	Calculated	Cosmic	87
Cent x Sub	Full	Cosmic	87
Full	Cent x Sub	Cosmic	86
Full	Cent Only	Cosmic	84
Calculated	Calculated	Cosmic	50

Figure 7.53 shows a plot of the total residual factors against the number of parameters. It can be seen that all but one of the residual factors reside in the range 50 to 85. When compared with the other results the final optimisation is comparable with the middle ranking parameter options and only 28% larger than the best case with 240% more optimising parameters. This comparison is slightly misleading as the final optimisation took place with the limited torsion angle option in place so a better comparison will have to be left to the comparison of the standard deviations of the individual interaction types. The situation that results in by far the worst result is the case where the strain free angle value is calculated using the 'central times substitution factor' option. In the following plots, where there is noticeable difference between the options, it can be seen that this is consistently the worst case.

Figure 7.54 shows the scatter plot for the standard deviations relating to the distance residual factors. As well as the points relating to the individual cases the plot also shows a line of best fit. This line shows a very shallow gradient which indicates a very limited trend for the standard deviation to decrease with parameter numbers. Compared with the variation in the standard deviations for the other structural properties this effect is negligible. This close similarity between all the optimisations is also shown in the small range of values for each bond length. In that, over all the bond lengths for all the optimisations, the average standard deviation is less than 0.003Å. So there are no problems relating to the distance properties caused by the combination of the reduced force field options.

Figure 7.55 shows the distribution of the standard deviations for the valence angle structural residual factors against the number of parameters. In this case each point is labelled to show the type of options used in the force field to determine the parameters for the valence angle interaction. The graph also shows the general trend of the residual factor when the worst case situation is ignored.

Figure 7.53 Plot of Residual Factors against Parameter Numbers

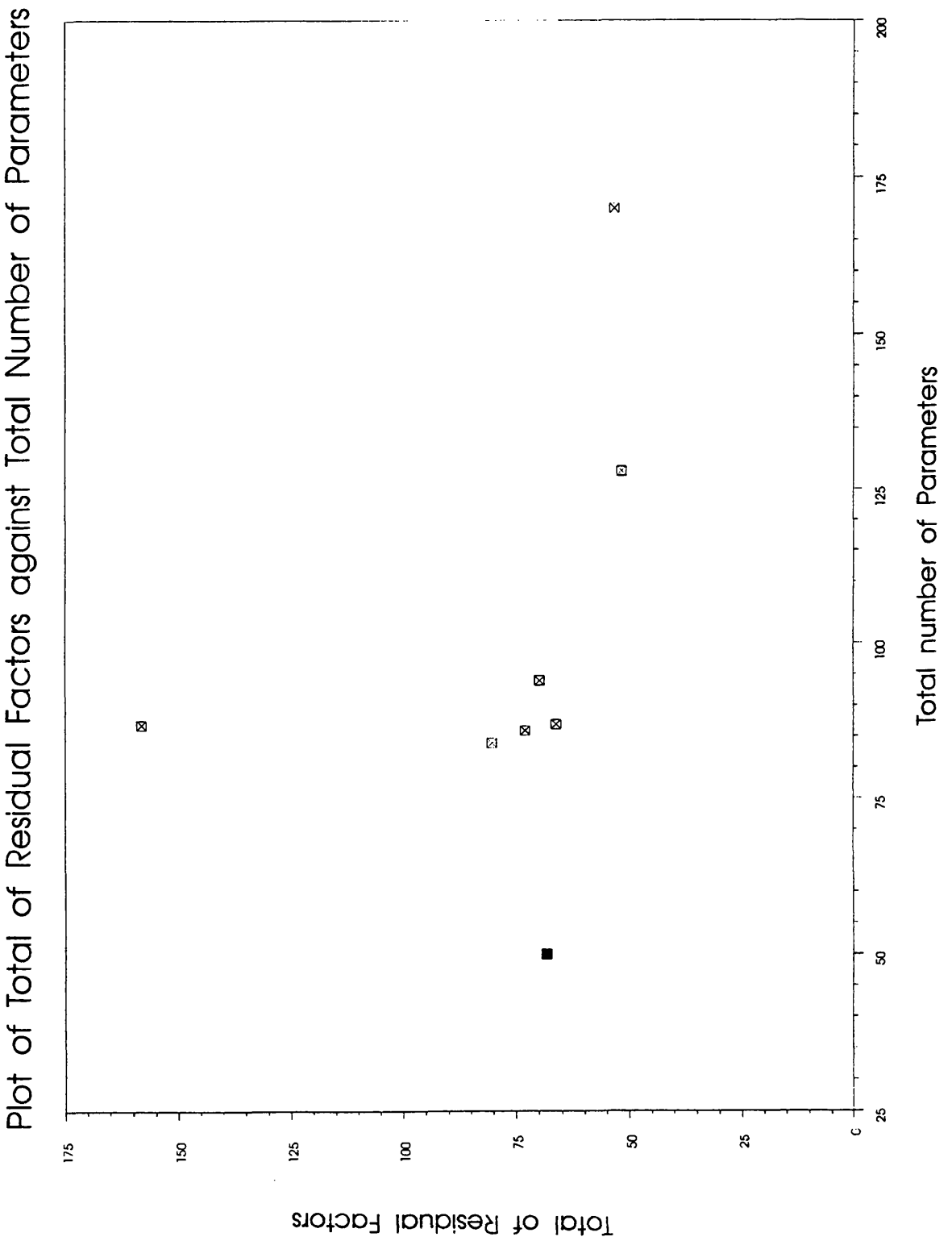


Figure 7.54 Plot of Distance S.D's against Parameter Numbers

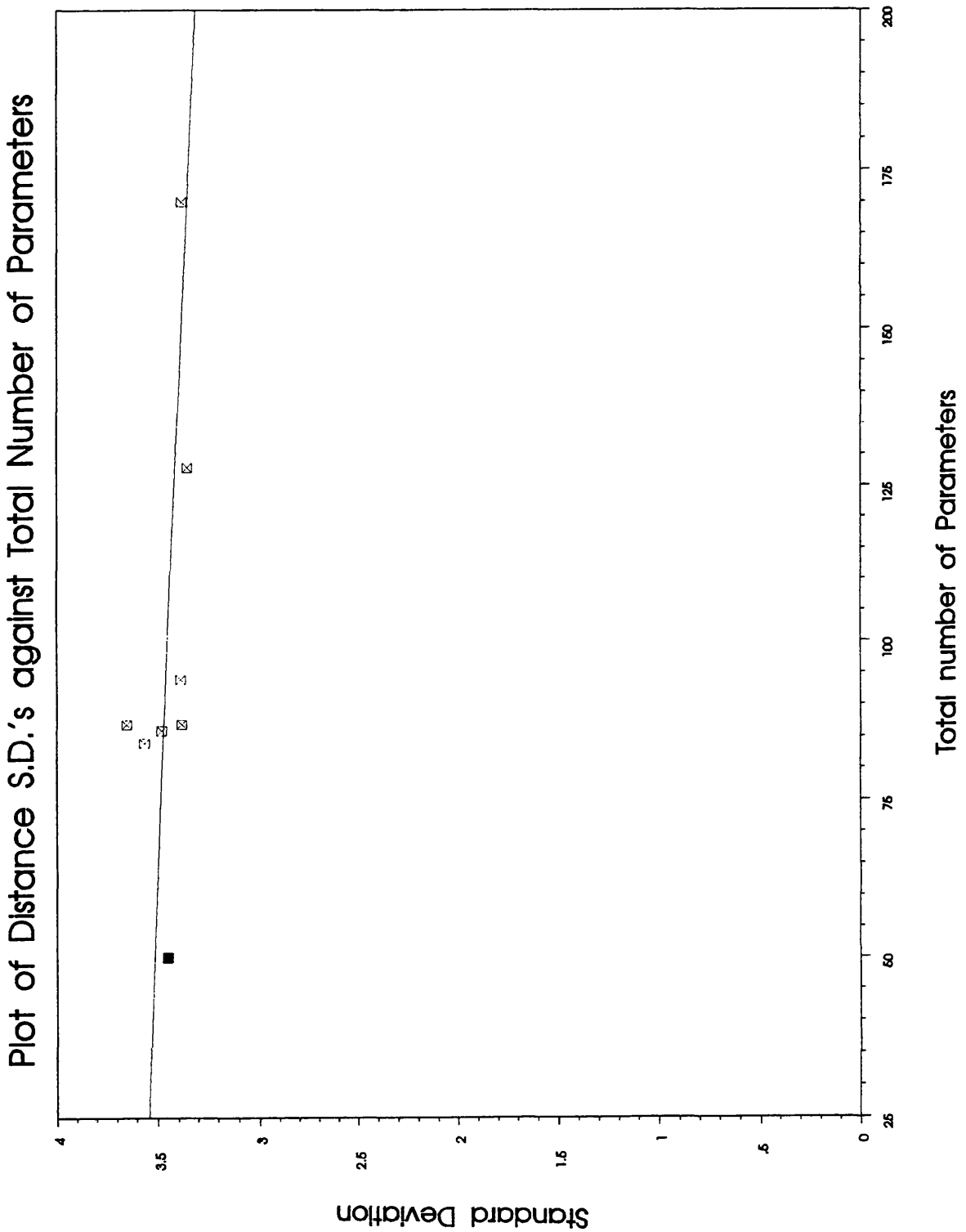
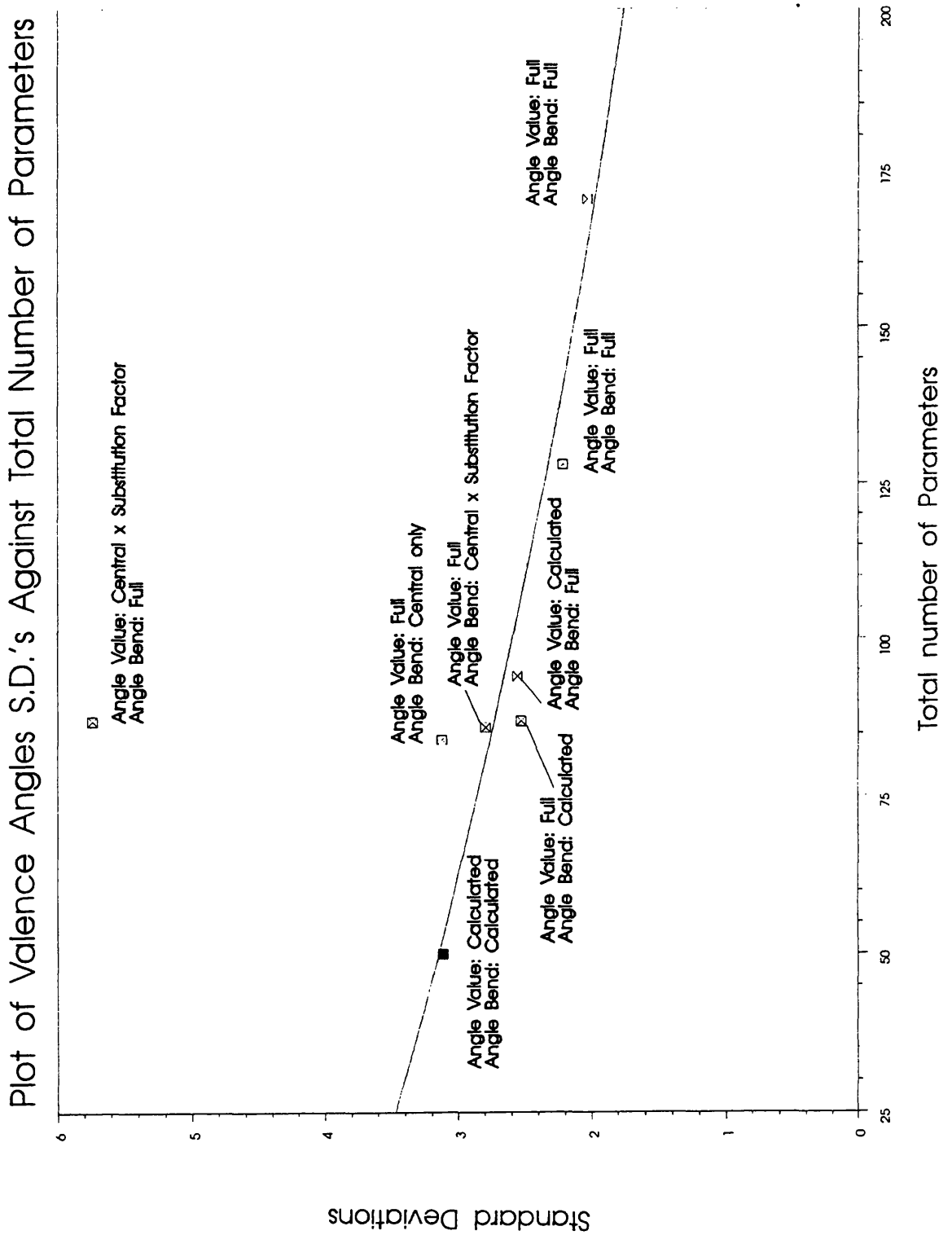


Figure 7.55 Plot of Valence Angle S.D.'s against Parameter Numbers



This graph shows how important it is for correct structural modelling to get the right option for the strain free valence angles as compared to the angle bending force constants. Examining the plot it can be seen that the case using the 'central times substitution factor' option for the angle value is twice as bad as when this type of option is used to determine the angle bending force constant, with the other option being 'full' in both cases.

This shows how sensitive the valence angle standard deviation residual factor is to the strain free angle option so it is encouraging to note that the point for the 'calculated' angle with the same 'full' option for the angle bend parameters is below the line showing the general trend.

This graph shows, probably better than was discussed earlier using figure 7.26, why the 'calculated' technique was chosen as the optimum angle bend option. The comparable points to look at in this case are the four points with total parameter numbers up to 128 that have their strain free angle values determined by the 'full' option. The first three points form an almost vertical line with a far greater gradient than the general trend. After the last of these, the 'calculated' option, the gradient of the line to the next point, the the 'full' option, is less then the general trend. This then shows categorically that the 'calculated' option is the best of the reduced options and that it compares favourably with full option. What has then to be remembered is that the choice of the optimum options took into account the number of parameters that would be required for a force field using more atom types and that as such the number of parameters that would be required in each of these cases are as follows:

Full	≈2000
Calculated	5

This enormous difference between the number of parameters taken with the closeness of the standard deviations and the earlier point about the angle bending force constants being less important than the strain free angles makes the 'calculated' option the superior choice for use in the final force field configuration.

Examining the result from the final optimisation that uses 'calculated' options for both it can be seen that, as would be expected, the result is worse than the cases where only one of the options was 'calculated', which are again worse than the comparable case where both the options were 'full'. If a best fit straight line is calculated for the three comparable situations involving one or more 'full' options it results in a predicted standard deviation of 2.880 when the total number of parameters is 50. This compares with the actual standard deviation of the final optimisation case of 3.119. This shows that the increase is only slightly greater than would be expected if the increase in the standard deviation was linear with the total

number of parameters. Table 7.11 shows the number of parameters that would be required for each of these combinations for a force field involving 40 atom types.

Table 7.11 Number of Parameters for Valence Angle Options

Both Calculated	52
One Calculated, One Full	≈2000
Both Full	≈4000

Thus, if the results for the restricted structure set used in this investigation can be extrapolated to the fuller case then it can be seen that the disadvantage of the fully calculated option being slightly less accurate is considerably outweighed by the advantages that are realised from the use of such a small number of parameters.

Figure 7.56 shows the standard deviations for the torsion angles. The problem with interpreting this graph is similar to the one that resulted in the inability to properly compare the torsion angle options with the selected weighting. That is the propensity of torsion angles to fall into the 'incorrect' minima making the interpretation of the advantages of different cases difficult.

If the RESULTS files for the given examples are examined and the torsion angles residual factor extracted then table 7.12 is obtained. This table contains the torsion angles in the order that they appear in the earlier list of structural properties for the final optimisation results. To help understand this table those entries that show a torsion angle residual factor of 4 or greater are highlighted.

It can be seen that several of the torsion angles have a tendency to have high residual factors. These fall into three categories.

A) Torsion angles residing in the correct minima where the experimental error is low. For example, in Gauche 1,1,2,2-Tetrafluoro Ethane, the sixth residual factor in the list, the experimental value for the torsion angle H-C-C-H is 78.0(2.0)° whereas the initial optimisations result in an angle of approximately 64°.

B) Torsion angles where the required minima is not present, either due to a deficiency in the torsional barriers or the non-bonded interactions. This appears to happen often around C_{sp2}-C_{sp2} bonds where the optimisation often results in a value of 180° where the experimental value indicates that the angle is not planar, for example, 155.3(5.5)° in Anti-Syn Fumaryl Chloride.

C) Torsion angles that end up in the wrong minima. Why the torsion angle should move into a different minimum is difficult to explain. It may be that the original structure starts off in the wrong configuration or that the optimising program somehow forces the structure over an energy barrier.

Figure 7.56 Plot of Torsion S.D's against Parameter Numbers

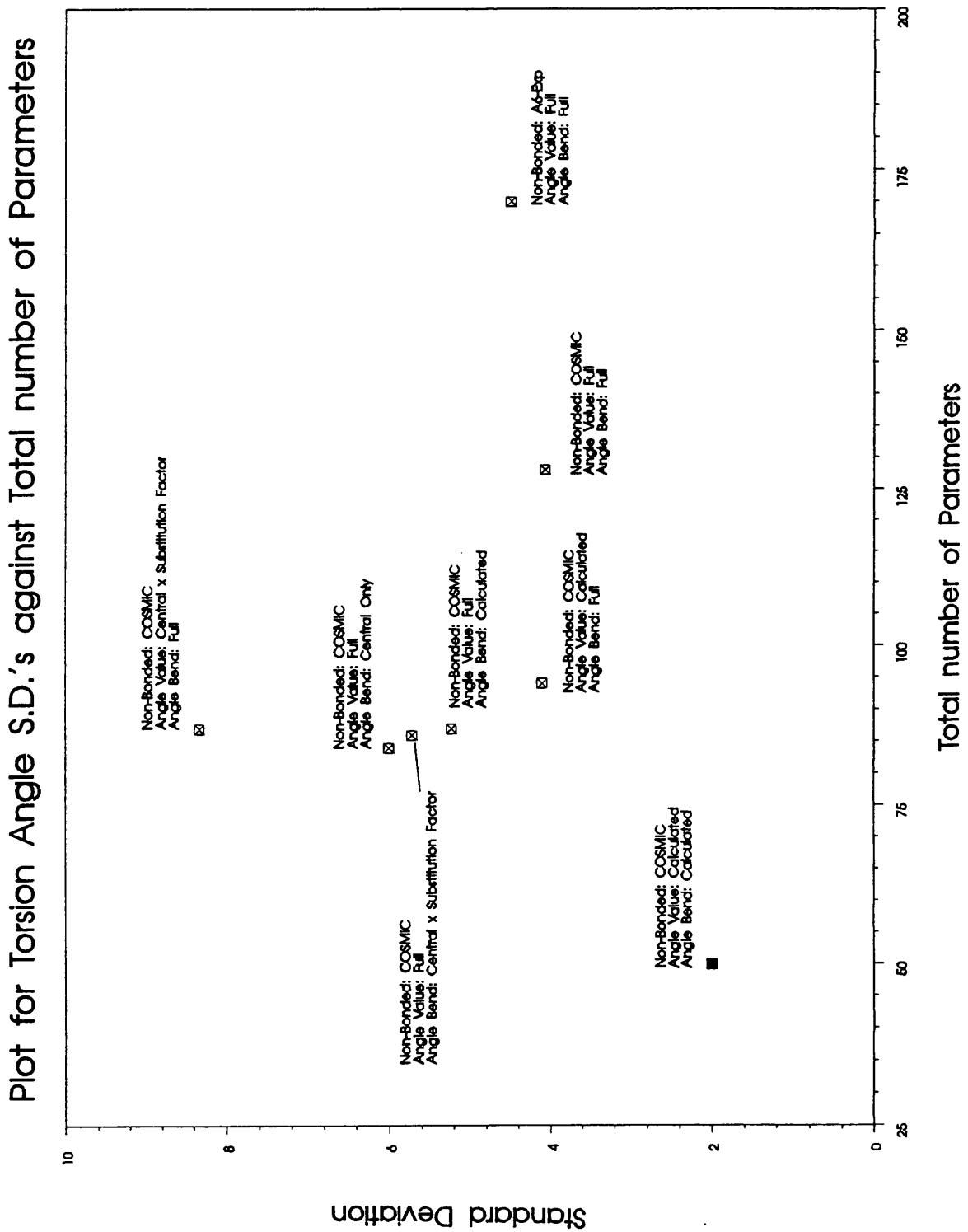


Table 7.12 Torsion Residual Factors for Optimisations

Non-Bonded Angle Value Angle Bend	Cosmic Calculated Calculated Parameters: 50	Cosmic Full Central Only Parameters: 84	Cosmic Full Cent x Sub Parameters: 86	Cosmic Full Calculated Parameters: 87	Cosmic Cent x Sub Factor Full Parameters: 87	Cosmic Calculated Full Parameters: 94	Cosmic Full Full Parameters: 128	Ab-Exp Full Full Parameters: 170
	0.0	0.0	0.0	0.0	15.12943	0.0	0.0	0.0
	0.21342	0.21519	0.21519	0.21519	0.69060	0.37019	0.0	0.0
	3.48582	3.49450	3.48731	3.49091	10.63406	3.60212	3.33591	3.18918
	0.01807	0.0	0.0	0.0	5.68692	0.0	0.00264	0.0
	0.0	0.0	0.0	0.0	7.25989	0.0	0.0	0.0
	4.134454	5.972610	5.972991	5.972610	5.91961	5.21898	5.972565	5.972610
	0.96326	0.78784	0.60368	0.39661	0.67443	0.79141	0.02840	0.33764
	0.0	23.774925	23.223203	23.153154	19.71013	19.21484	21.793669	21.779957
	0.49129	0.82154	0.76726	0.56339	0.81444	0.52072	0.0	0.59940
	1.12723	1.25494	1.42660	1.53492	1.55259	1.51319	0.47031	1.58605
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.23157	24.105763	24.225961	21.977196	1.13822	0.0	0.0	1.77420
	4.013810	4.437457	4.18229	1.65725	0.10544	0.31222	10.930908	1.05155
	4.009930	6.972210	6.83995	6.63813	5.96372	6.54102	6.026972	6.023743
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.61914	0.15883	0.83593	0.0	0.0	0.0	0.25468
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	4.155161	2.36328	6.427646	0.42953	0.0	0.0	1.47634	0.12149
	0.08949	1.07191	0.60088	0.05831	0.0	7.01217	5.338857	0.0
	0.0	0.0	0.0	1.20427	0.0	0.0	0.0	1.07008
	1.94399	0.60167	0.78357	0.61354	0.0	0.0	0.06228	0.0
	0.0	0.0	0.0	0.0	0.61921	1.41123	0.70664	0.89625
	1.31456	0.0	0.0	0.0	2.14022	0.0	0.0	0.0
	0.08916	0.0	0.26961	0.59777	0.0	0.18610	0.0	0.0
	0.0	0.0	0.0	0.0	0.42941	0.0	0.64679	0.02016
	0.0	0.0	0.0	0.0	2.16430	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1.71396	1.85634	2.09314	2.19164	6.31620	2.44873	3.53342	2.71472
	0.0	0.0	0.0	0.0	27.51321	0.0	0.0	0.0
	3.99491	6.024419	3.99713	3.47435	8.09960	4.70004	2.22173	5.940336
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	4.010451	2.75579	2.33505	2.50148	0.48248	3.23074	1.37803	1.93828
	0.84533	0.55640	0.02663	0.0	1.78786	0.0	1.03830	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1.89559	1.56735	1.79816	2.00407	1.01806	1.62693	1.62716	2.10353
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	4.017745	11.311863	12.119644	13.112177	4.40195	13.49100	16.224231	11.332233
	0.0	0.0	0.0	0.0	5.19448	0.0	0.0	0.0
	3.17736	2.89029	2.75480	2.66190	5.45781	3.39974	1.38343	2.74854
	1.64909	0.24619	0.42505	0.75475	9.77081	0.86330	0.0	0.0
	2.58087	1.68903	1.71151	2.05643	14.89162	1.91988	0.11297	0.0
	4.008881	17.995688	15.867652	8.166609	7.53495	7.46335	0.0	9.330725
	4.005519	9.409835	0.81830	2.29578	9.81938	4.38608	2.02392	0.0
	3.60429	3.73383	3.00032	3.21474	5.17285	3.54346	5.06144	3.86974

Looking at the COMPARE file for the optimisation using 128 parameters, which is used in most of the comparisons, it is found that of the 7 torsion angles with residual factors of 4.0 or more, 4 are of type A, 2 of type B and 1 of type C. A similar investigation for the optimisation with the 'full(trial)' option shows the same numbers for type A and B but has two more high residual factors that appear to be of type C.

Because of these problems there is little that can be deduced from this graph. What it does show is that the few cases that are limited in the final optimisation have a large effect on the standard deviation. This can be seen from the fact that the standard deviation for the final optimisation is half that of any of the previous optimisations with the same weighting. It was seen earlier that when the weighting was even between residual factors that this tendency of the torsion angle to be grossly different was reduced. It may be that the weighting values resulted in the excessive favouring of the energy differences over the structural properties.

Figure 7.57 shows the distribution of the standard deviations for the energy differences against parameter numbers. Some features of this graph are:

a) The worst situation, as in all the other cases, is the optimisation which used the 'Central times Substitution Factor' option to calculate the strain free angle.
b) The best case is that which uses the greatest number of parameters.
c) It can be seen that, as with the previous cases, there is a tendency for the residual factor to decrease with an increase in the number of parameters. It can be seen that the final case is lower than would be expected if this trend was extrapolated from the earlier optimisations. It has, however, to be remembered that this final optimisation had the limiting option on the torsion angles and that this had a significant effect on one of the structures involved in the energy difference calculations.

Figure 7.58 shows the distribution of the standard deviations for the heat of formation residual factors. As shown in the earlier sections the heat of formation residual factor is the measure most improved by using a large number of parameters. The plot can be divided into three sections depending upon the number of angle parameters. The lower section contains the two cases where both the angle parameters' used the 'full' option. The next section relates to those cases where the strain free angle value was determined in the 'full' manner. In this case the three options in this section are very close.

The third section relates to those options that have a reduced option for the strain free angle. Of these it can be seen that both cases with the 'calculated' option is again better than the 'central times substitution' case.

Figure 7.57 Plot of Energy Dif S.D's against Parameter Numbers

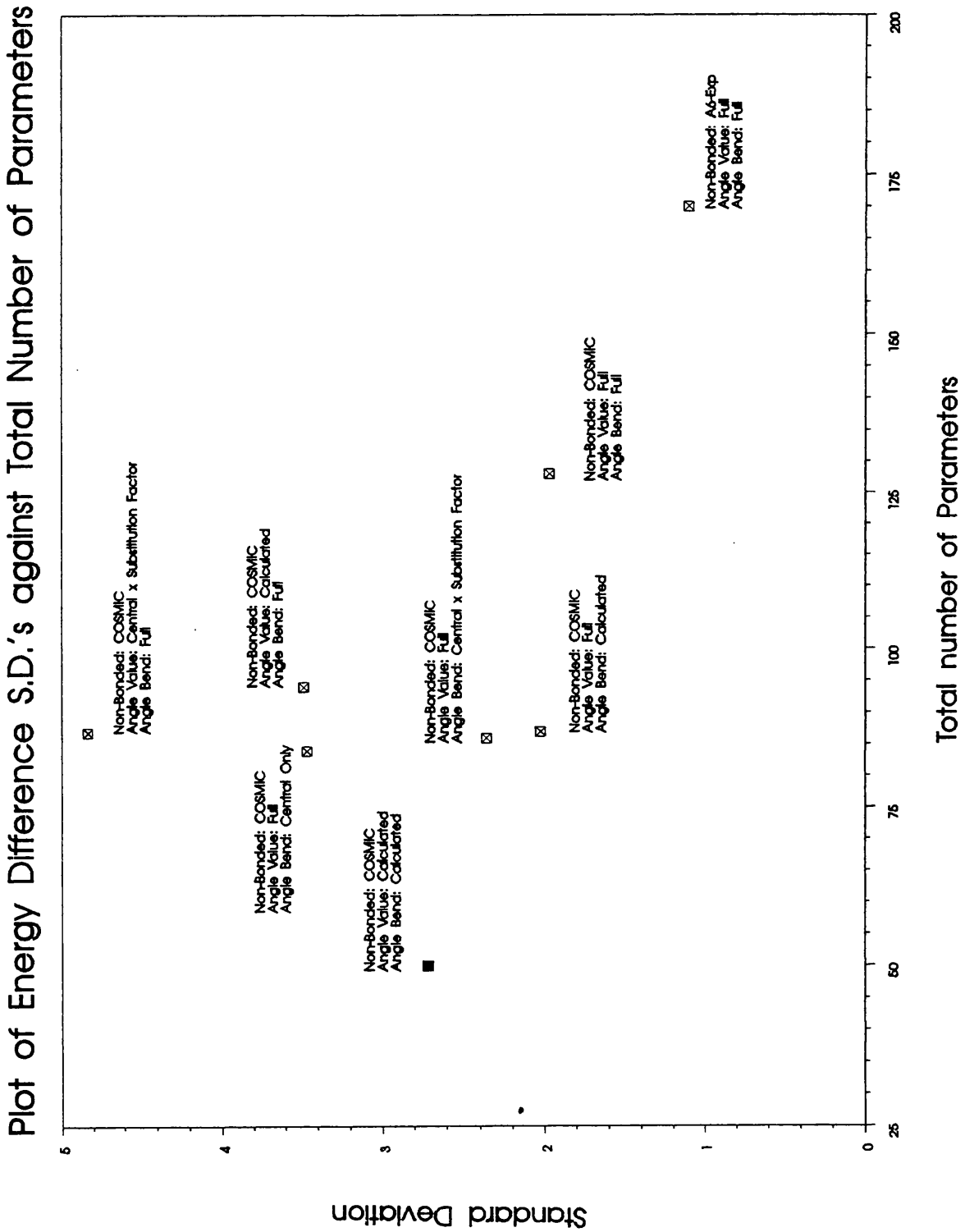
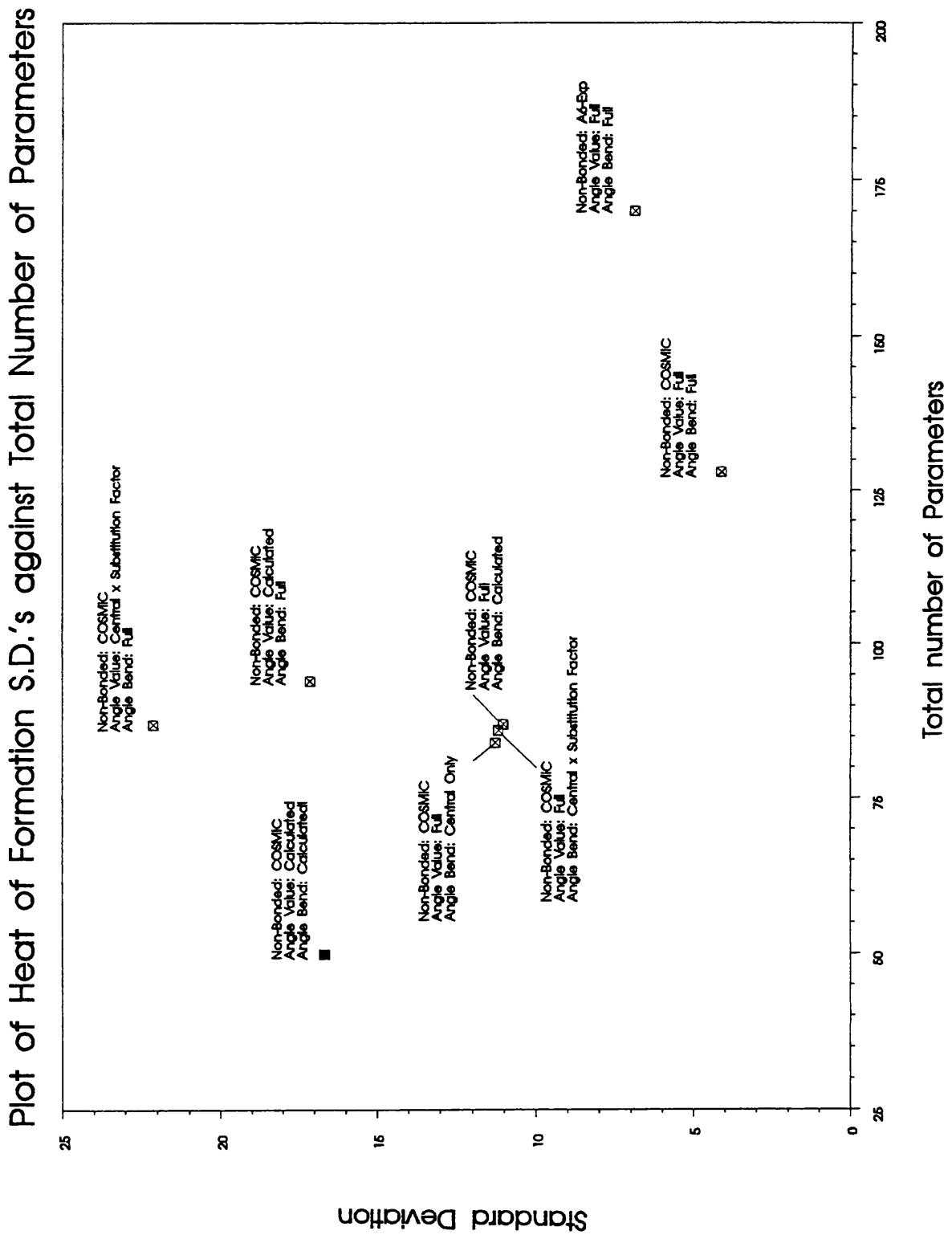


Figure 7.58 Plot of Heat of Formation S.D's against Parameter Numbers



Overall the graphs show that the highly reduced force field used in final optimisation compares well with the other combinations. In particular the use of the 'calculated' option for determining the strain free valence angles can be seen in figure 7.55 to be far superior to the other reduced method of determining these parameters and to be close to the 'full' method. Figure 7.55 also shows that for the angle bending parameters the combination of reduced parameter options is effectively linear with the number of parameters. That is the increase in the residual factors appears to follow a relatively linear path rather than some exponential curve. This is important as it shows that the reduced parameter options can apparently be combined such that the combination does not lead to any loss in accuracy other than which would be expected from the extrapolation of cases where not all the options were reduced.

7.5.2 Final Force Field Parameters

The following are the parameters resulting from the final force field optimisation, where possible they are compared with similar parameters from other force fields. Note that all the optimisable parameters have been given to 6 decimal places. Investigation showed that, as might be expected, rounding parameters to 4 or 5 significant figures resulted in different residual factors. Thus to ensure consistency with the residual factors all parameters are shown to the maximum number of significant figures given in the saved parameter files.

7.5.2.1 Bond Lengths

Table 7.13 Optimised Bond Lengths from Final Optimisation

Bond	This Study	Cosmic
H-Csp2	1.085964Å	1.089Å
H-Csp3	1.095005Å	1.100Å
Csp2=Csp2	1.324988Å	1.335Å
Csp2-Csp3	1.476049Å	1.501Å
Csp3-Csp3	1.487227Å	1.520Å
Cps2=O	1.196046Å	1.220Å
Csp2-F	1.327037Å	1.330Å
Csp3-F	1.340696Å	1.360Å
Csp2-Cl	1.742544Å	1.750Å
Csp3-Cl	1.766968Å	1.767Å
Csp2-Br	1.911403Å	------
Csp3-Br	1.923321Å	1.938Å
Csp2-Csp2	1.445262Å	------

As would be expected there are few remarkable points about the bond length values. A couple of values that come out considerably shorter than those in the other force field are all the Carbon-Carbon distances that are 0.022Å shorter on average in this study. One reason for this may be that the values from the COSMIC force field for these bonds come from the WBFF where the study was on a hydrocarbons only structure set whereas this study involves other atom types.

7.5.2.2 Bond Stretching Force Constants

The following are the optimised parameters used to determine the bond stretch force constants.

Table 7.14 Optimised Parameters for Calculating Bond Stretch Force Constants

c1	1.000000Å	(fixed)
c2	1524.469849 Kcal Mol^{-1}	
c3	4.459719 Kcal Mol^{-1}	
c4	791.911255 Kcal Mol^{-1} $Å^{-1}$	
c5	287.974091 Kcal Mol^{-1}	

These are used with the following equations to produce the force constants in table below:

$$½k_l(ij) = \left[\frac{c_2}{l_0(ij)^2}\right] + \left[\frac{c_3}{(l_0(ij) - c_1)^2}\right] - \left[\frac{c_4}{l_0(ij)}\right]$$

$$½k_l(iH) = \frac{c_5}{l_0(iH)^2}$$

Table 7.15 Calculated Force Constants for the Final Optimisation

Bond	This Study	Cosmic	CHEMMIN
All values are in Kcal Mol^{-1} $Å^{-2}$			
H-Csp2	244.2	346.0	334.9
H-Csp3	240.2	331.2	329.4
Csp2=Csp2	312.9	670.0	594.4
Csp2-Csp3	182.9	319.5	407.1
Csp3-Csp3	175.5	316.8	396.6
Cps2=O	519.9	777.6	903.6
Csp2-F	310.6	600.0	591.1
Csp3-F	295.9	300.0	569.7
Csp2-Cl	55.7	260.0	222.6
Csp3-Cl	47.7	300.0	201.8
Csp2-Br	8.3	-----	151.7
Csp3-Br	5.5	300.0	147.5
Csp2-Csp2	204.4	-----	438.0

Table 7.15 compares the bond stretch force constants determined using the final parameter values using the previously given values of the bond lengths with the appropriate force constants in COSMIC and with the same equation but with the CHEMMIN parameters (c1=1.0, c2=1800.0, c3=7.90, c4=670.0, c5=395.0).

It can be seen that the values from this study are consistently lower than those determined by other methods.

This must be seen as a set of parameters where, with the current structure set at least, the parameters cannot be optimised in the purely automatic manner that it is done within OPTPARAM.

It should though be noted that the values calculated with the parameters used in the CHEMMIN program are far more realistic. Thus it should be a simple task to optimise a more realistic force field by simple optimising again but with these parameters fixed to the CHEMMIN values. Due to time constraints this was not done, but it can be expected that this would not have a significant effect on the overall results as the deviation of most bonds from their strain free lengths is small.

This is reinforced by the fact that at the beginning of the final optimisation the bond stretch force constant parameters were reset to the values used in CHEMMIN and that during the final optimisation the change in the average bond distance error was only 0.003Å.

In conclusion these parameters are difficult to optimise with the current structure set but reasonable values are available elsewhere which, if used instead are unlikely to result in a large distortion in the final results.

7.5.2.3 Non-Bonded Parameters

The table 7.16 gives the values for the non-bonded parameters at the end of the final optimisation. These values relate to the COSMIC non-bonded interaction equation.

$$E_{VdW} = f\left[\frac{-2.25}{rr^6} + 8.28\times10^5 \cdot e^{\left(\frac{-rr}{0.0736}\right)}\right]$$

Where rr is the distance between the atoms divided by the sum of the van der Waals radii and f is the product of two 'well depth' parameters relation to the atom types involved, i.e.:

$$f_{CN} = f_C \times f_N$$

Table 7.16 Optimised Non-Bonded Parameters

Non-Bonded Parameters		This Study	Cosmic
VdW Radius	H	1.528836Å	1.20Å
VdW Radius	Csp2	1.608123Å	1.70Å
VdW Radius	Csp3	1.808644Å	1.70Å
VdW Radius	O	1.981703Å	1.52Å
VdW Radius	F	2.037748Å	1.47Å
VdW Radius	Cl	1.540521Å	1.75Å
VdW Radius	Br	1.641148Å	1.85Å
f	H	0.095945	0.042
f	Csp2	0.436700	0.107
f	Csp3	0.086834	0.107
f	O	0.052127	0.116
f	F	0.005076	0.109
f	Cl	0.169202	0.314
f	Br	0.131224	0.434

This another situation where some of the values produced by the optimisation do not correlate with what might be expected. Although it would be good for these values to be close it has to be remembered that molecular mechanics is empirical in nature and the difference in parameters between force fields show that there is no necessity for a parameter to have some universal value.

A probable reason why some of the above parameters, such as the van der Waals radii, are different from accepted values is that there are few examples in the structure set of distances between non-bonded atoms. This lack of relevant structural properties is compounded by the fact that non-bonded interactions are only taken for 1-4 or greater interactions, that is atoms that cannot be connected on a common atom. The result of this is that they most affect the torsion barriers where a balance has to be obtained between a torsional minimum position and a non-bonded close contact. There are unfortunately only a few of these torsion angles in the structure set and the generalised nature of the torsion barrier means that these few cases are ripe for the production of artefacts rather than realistic parameters.

7.5.2.4 Strain Free Angles

Angle Value

Table 7.17 contains the optimised parameters used to calculate the strain free valence angles. As stated each angle about the central atom is given a weighting factor:

$$\text{weighting} = (S_i + S_k) / (l_0(ij)^a \times l_0(jk)^a)$$

Where S_i and S_k are size factors for atoms i and k respectivly, $l_0(ij)$ and $l_0(jk)$ are the associated bond lengths and a is another optimisable parameter. These weighting factors are then used to divide up the total angle associated with the central atom type, i.e.:

$$\theta_n = \text{Total Angle} \times \frac{\text{Weighting for angle n}}{\text{Total of all weightings}}$$

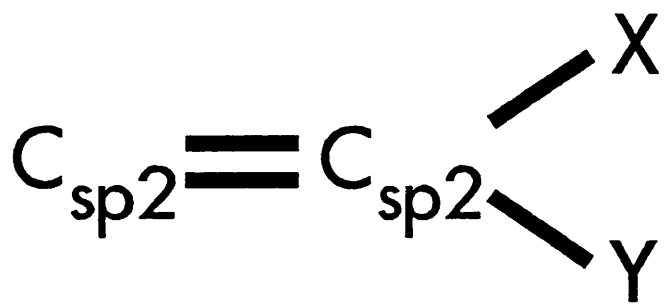
Table 7.17 Optimised Parameters Used to Calculate Strain Free Angles
All atom weights are unitless

Size Factor	H	106.178909
Size Factor	Csp2	136.637939
Size Factor	Csp3	133.968964
Size Factor	O	145.650192
Size Factor	F	118.442848
Size Factor	Cl	154.256982
Size Factor	Br	167.572372
a (Exponential)		0.355492
Cps2 Total Angle		360.0° (fixed)
Csp3 Total Angle		657.0° (fixed)

It is difficult to compare the strain free angles produced by this method with most force fields as few tend to give strain free angles depending on both the central atom type and all the atoms bonded to it. All the calculated angles for a central C_{sp3} atom are given in appendix K.

The following two pages show all the angle combinations using a central C_{sp2} atom, followed by an extract from the list of possible combinations for C_{sp3}.

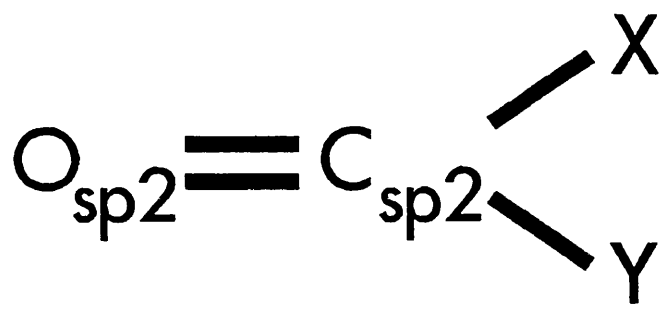
Table 7.18 All Possible Combinations for $C_{sp2}=C_{sp2}$



Y	H									
H	1) 122.5 2) 122.5 3) 115.0									
	Csp2									
Csp2	1) 120.55 2) 122.55 3) 117.0	1) 121.25 2) 121.25 3) 117.55								
		Csp3								
Csp3	1) 121.95 2) 121.9 3) 116.1	1) 122.6 2) 120.55 3) 116.85	1) 121.9 2) 121.9 3) 116.2					X		
			F							
F	1) 124.0 2) 121.3 3) 114.65	1) 124.35 2) 119.65 3) 116.0	1) 123.65 2) 121.05 3) 115.3	1) 122.95 2) 122.95 3) 114.1						
					Cl					
Cl	1) 120.6 2) 122.1 3) 117.3	1) 121.6 2) 121.05 3) 117.3	1) 120.85 2) 122.4 3) 116.75	1) 119.85 2) 124.0 3) 116.15	1) 121.55 2) 121.55 3) 116.9					
								Br		
Br	1) 119.45 2) 122.4 3) 118.2	1) 120.6 2) 121.55 3) 117.85	1) 119.85 2) 122.9 3) 117.25	1) 118.75 2) 124.4 3) 116.9	1) 120.65 2) 122.15 3) 117.2	1) 121.35 2) 121.35 3) 117.35				

Angle 1 Csp2=Csp2-X
Angle 2 Csp2=Csp2-Y
Angle 3 X-Csp2-Y

Table 7.19 All Possible Combinations for $C_{sp2}=O_{ps2}$



Y	H	Csp2	Csp3	X		
H	1) 125.3 2) 125.3 3) 109.35					
Csp2	1) 123.55 2) 115.1 3) 111.35	1) 123.9 2) 123.9 3) 112.15				
Csp3	1) 124.95 2) 124.45 3) 110.6	1) 125.3 2) 123.2 3) 111.5	1) 124.6 2) 124.6 3) 110.8			
F	1) 126.95 2) 123.95 3) 109.1	1) 126.95 2) 122.45 3) 110.6	1) 126.3 2) 123.85 3) 109.85	1) 125.7 2) 125.7 3) 108.65		
Cl	1) 123.7 2) 124.45 3) 111.85	1) 124.35 2) 123.6 3) 112.05	1) 123.65 2) 124.95 3) 111.45	1) 122.7 2) 126.45 3) 110.8	1) 124.15 2) 124.15 3) 111.7	
Br	1) 122.55 2) 124.65 3) 112.75	1) 123.4 2) 124.0 3) 112.6	1) 122.7 2) 125.35 3) 112.0	1) 121.7 2) 126.75 3) 111.55	1) 123.3 2) 124.65 3) 112.0	1) 123.9 2) 123.9 3) 112.2

Angle 1 Osp2=Csp2-X
Angle 2 Osp2=Csp2-Y
Angle 3 X-Csp2-Y

Table 7.20 Selection of possible Combinations for C_{sp3}

Atoms	X	X	X	X
X	X	Angle =	109.500	

Atoms	H	Csp2	Csp3	F
H	Csp2	Angle =	111.645	
H	Csp3	Angle =	110.095	
H	F	Angle =	106.845	
Csp2	Csp3	Angle =	111.592	
Csp2	F	Angle =	109.140	
Csp3	F	Angle =	107.683	

Atoms	H	Csp2	Csp3	Cl
H	Csp2	Angle =	109.510	
H	Csp3	Angle =	107.990	
H	Cl	Angle =	110.111	
Csp2	Csp3	Angle =	109.458	
Csp2	Cl	Angle =	110.633	
Csp3	Cl	Angle =	109.298	

Atoms	H	Csp2	Csp3	Br
H	Csp2	Angle =	108.585	
H	Csp3	Angle =	107.078	
H	Br	Angle =	111.398	
Csp2	Csp3	Angle =	108.533	
Csp2	Br	Angle =	111.352	
Csp3	Br	Angle =	110.053	

Atoms	H	Csp2	Cl	Br
H	Csp2	Angle =	107.984	
H	Cl	Angle =	108.576	
H	Br	Angle =	110.781	
Csp2	Cl	Angle =	109.091	
Csp2	Br	Angle =	110.736	
Cl	Br	Angle =	109.831	

Atoms	H	Csp3	F	Cl
H	Csp3	Angle =	109.979	
H	F	Angle =	106.733	
H	Cl	Angle =	112.139	
Csp3	F	Angle =	107.569	
Csp3	Cl	Angle =	111.311	
F	Cl	Angle =	109.269	

Atoms	H	Csp3	Cl	Br
H	Csp3	Angle =	107.155	
H	Cl	Angle =	109.259	
H	Br	Angle =	111.478	
Csp3	Cl	Angle =	108.453	
Csp3	Br	Angle =	110.132	
Cl	Br	Angle =	110.522	

Atoms	H	Csp3	F	Br
H	Csp3	Angle =	108.988	
H	F	Angle =	105.771	
H	Br	Angle =	113.385	
Csp3	F	Angle =	106.600	
Csp3	Br	Angle =	112.016	
F	Br	Angle =	110.239	

Atoms	H	F	Cl	Br
H	F	Angle =	105.096	
H	Cl	Angle =	110.419	
H	Br	Angle =	112.662	
F	Cl	Angle =	107.593	
F	Br	Angle =	109.536	
Cl	Br	Angle =	111.695	

Atoms	Csp2	Csp3	F	Cl
Csp2	Csp3	Angle =	110.746	
Csp2	F	Angle =	108.313	
Csp2	Cl	Angle =	111.935	
Csp3	F	Angle =	106.867	
Csp3	Cl	Angle =	110.584	
F	Cl	Angle =	108.555	

Atoms	Csp2	F	Cl	Br
Csp2	F	Angle =	106.908	
Csp2	Cl	Angle =	110.483	
Csp2	Br	Angle =	112.148	
F	Cl	Angle =	107.147	
F	Br	Angle =	109.082	
Cl	Br	Angle =	111.232	

Atoms	Csp3	F	Cl	Br
Csp3	F	Angle =	106.138	
Csp3	Cl	Angle =	109.830	
Csp3	Br	Angle =	111.531	
F	Cl	Angle =	107.815	
F	Br	Angle =	109.762	
Cl	Br	Angle =	111.926	

7.5.2.5 Valence Angle Bend Force Constants

The equations for the calculations of the valence angle bending force constants are as follows.

$$½k_\theta(ijk) = \frac{d_1\left\{d_2 + d_3\left(|X(i) - X(j)| + |X(j) - X(k)|\right)\right\}}{l_0(ij)l_0(jk)}$$

$$½k_\theta(ijH) = \frac{d_4 \times d_1\left\{d_2 + d_3\left(|X(i) - X(j)| + |X(j) - X(H)|\right)\right\}}{l_0(ij)l_0(jH)}$$

$$½k_\theta(HjH) = \frac{d_5 \times d_1\left\{d_2 + d_3\left(|X(H) - X(j)| + |X(j) - X(H)|\right)\right\}}{l_0(Hj)l_0(jH)}$$

Where d_1-d_5 are parameters that can be optimised in the same way as c_1-c_4. X(i) is the electronegativity of the atom i. The following section contains the results from the final optimisation and lists the equivalent parameters from the CHEMMIN program.

Table 7.21 Optimised Parameters for Calculating Angle Bending Force Constants

Angle Bend	This Study	CHEMMIN
d1	0.001146	0.001388 Kcal Mol^{-1} Deg^{-2} $Å^2$
d2	20.075312	15.0
d3	0.484955	2.33
d4	0.857282	0.45
d5	0.997448	0.20

Again, as this is a reduced parameter option where the final parameters are calculated using the above parameters, it is difficult initially to determine if these parameters are reasonable. One problem that can be spotted straight away is the large values of the scaling factors (d4 and d5) for angles including 1 or 2 hydrogens. The main reason for this is almost certainly that the electron density studies report few angles involving hydrogens and those given have large errors. A direct comparison between the new results and those from the CHEMMIN program can be seen in the following table of calculated valence angle bend force constants where the first column contains values calculated for this study, the second values calculated using the parameters used in the CHEMMIN and the third gives the percentage comparison between the first two columns.

Cps2

```
Two Hydrogens

   H    -  Csp2 -  H       .01978      .00397      498.17
```

One Hydrogen

H	-	Csp2	=	Csp2	.01384	.00692	200.10
H	-	Csp2	-	Csp2	.01269	.00634	200.10
H	-	Csp2	-	Csp3	.01243	.00621	200.10
H	-	Csp2	=	Osp2	.01570	.00878	178.75
H	-	Csp2	-	F	.01431	.00842	169.99
H	-	Csp2	-	Cl	.01065	.00564	188.70
H	-	Csp2	-	Br	.00966	.00500	193.07

No Hydrogens

Csp2	=	Csp2	-	Csp2	.01201	.01087	110.50
Csp2	=	Csp2	-	Csp3	.01177	.01065	110.50
Csp2	=	Csp2	-	F	.01356	.01460	92.87
Csp2	=	Csp2	-	Cl	.01008	.00972	103.78
Csp2	-	Csp2	-	Br	.00915	.00860	106.35
Csp2	-	Csp2	-	Csp2	.01101	.00997	110.50
Csp2	-	Csp2	-	Csp3	.01079	.00977	110.50
Csp2	-	Csp2	=	Osp2	.01363	.01392	97.95
Csp2	-	Csp2	-	F	.01243	.01338	92.87
Csp2	-	Csp2	-	Cl	.00925	.00891	103.78
Csp2	-	Csp2	-	Br	.00839	.00789	106.35
Csp3	-	Csp2	-	Csp3	.01057	.00957	110.50
Csp3	-	Csp2	=	Osp2	.01336	.01363	97.95
Csp3	-	Csp2	-	F	.01218	.01311	92.87
Csp3	-	Csp2	-	Cl	.00906	.00873	103.78
Csp3	-	Csp2	-	Br	.00822	.00773	106.35
Osp2	=	Csp2	-	F	.01537	.01821	84.40
Osp2	=	Csp2	-	Cl	.01144	.01232	92.87
Osp2	=	Csp2	-	Br	.01038	.01095	94.82
F	-	Csp2	-	F	.01401	.01733	80.84
F	-	Csp2	-	Cl	.01043	.01180	88.38
F	-	Csp2	-	Br	.00946	.01050	90.11
Cl	-	Csp2	-	Cl	.00776	.00792	97.95
Cl	-	Csp2	-	Br	.00704	.00703	100.19
Br	-	Csp2	-	Br	.00639	.00623	102.55
Csp2	-	Csp2	-	Csp2	.01101	.00997	110.50

Csp3

Two Hydrogens

H	-	Csp3	-	H	.01945	.00390	498.17

One Hydrogen

H	-	Csp3	-	Csp2	.01233	.00616	200.10
H	-	Csp3	-	Csp3	.01223	.00611	200.10
H	-	Csp3	-	F	.01405	.00827	169.99
H	-	Csp3	-	Cl	.01041	.00552	188.70
H	-	Csp3	-	Br	.00952	.00493	193.07

No Hydrogens

Csp2	- Csp3 -	Csp2	.01057	.00957	110.50
Csp2	- Csp3 -	Csp3	.01049	.00949	110.50
Csp2	- Csp3 -	F	.01206	.01298	92.87
Csp2	- Csp3 -	Cl	.00893	.00861	103.78
Csp2	- Csp3 -	Br	.00817	.00768	106.35
Csp3	- Csp3 -	Csp3	.01040	.00941	110.50
Csp3	- Csp3 -	F	.01196	.01287	92.87
Csp3	- Csp3 -	Cl	.00886	.00854	103.78
Csp3	- Csp3 -	Br	.00810	.00762	106.35
F	- Csp3 -	F	.01373	.01698	80.84
F	- Csp3 -	Cl	.01018	.01152	88.38
F	- Csp3 -	Br	.00931	.01033	90.11
Cl	- Csp3 -	Cl	.00755	.00770	97.95
Cl	- Csp3 -	Br	.00690	.00689	100.19
Br	- Csp3 -	Br	.00631	.00615	102.55

These show that it is mainly the aforementioned Hydrogen scaling factors which result in any significant difference between these values. For the other values it can be seen that the change in the parameters has not greatly affected the calculated values with no values being more than 20% off from those calculated from the CHEMMIN parameters from which they were optimised from. It should also be noted that the values for these parameters before the final optimisation were close to the CHEMMIN values and that during the optimisation to the above values the fall in the residual factors was small indicating that the large difference in the calculated bending force constants before and after has not had a significant effect on the overall residual factor.

7.5.2.6 Torsional Barriers

The following is a table of the torsional barriers from the final optimisation and for the same properties from the COSMIC force field.

Table 7.22 Optimised Torsion Barrier Parameters

Torsional Barriers	This study	Cosmic
All units Kcal Mol^{-1}		
Csp2=Csp2	5.737006	6.0
Cps2-Cps3	0.086213	0.06
Csp3-Csp3	0.078062	0.10
Csp2-Csp2*	0.812628	----

These show that optimised values are close to those from another force field determined by normal methods. The main question is if such a large generalisation to a single torsional barrier between pairs of atom types with no reference to the other two atoms involved is a legitimate approach? This point is discussed further in the discussion section.

7.5.2.6 Out of Plane Bending

The following is the single parameter used in the out of plane bending calculation. It can be seen that even after all the force field loops in the final optimisation its value is still close to its initial value taken from the WBFF. That this parameter has optimised to a value capable of producing a significant contribution to the total energy indicates the requirement for such an energy contribution to the total.

Table 7.23 Optimised Out-of-Plane Parameter

Out of Plane Bending	This study	WBFF
Csp2	0.002475	0.002

7.6 Discussion

The main positive conclusion that can be deduced from these results is that, when attempting to model the structural and energetic properties of a structure set such as that given here, a highly reduced force field using calculated parameters can produce results comparable with those produced by force fields using more traditional methods of parameter determination.

There are, unfortunately, several negative conclusions that can be made. Main amongst these is that the force field that is produced is highly dependant on the structure set under investigation and for several of the parameter types the resultant parameters are significantly different from those determined by normal means. Of these the most striking are the bond stretch force constants and the non-bonded parameters.

For the bond stretch force constants this problem can be directly linked to the fact that for bonded interactions the bond length parameters are far more significant in determining the actual minimised bond lengths and that the bond stretching contributes little to the strain free energy. This combined with the lack of vibrational frequency information means that the bond stretch force constants are highly susceptible to distortion during automatic optimisation.

For the non-bonded options the problem is that there are only a few cases in the structure set where a combination of close atomic proximity and actually acquiring the relevant structural properties from the journal result in a non-bonded interaction having a significant effect on a residual factor. This problem means that the ratio of situations where a non-bonded parameter is significantly used against the number of parameters is small, a situation ripe for the production of spurious artefacts amongst the parameters.

These points do not discount any assumptions of the useability of the highly reduced force field. Indeed if the ratio of observed properties to optimisable

parameters is considered then it could be said that this gives a greater opportunity for the production of artefacts for the normal methods thus effectively reducing the real gap between the methods. There is though little point in this assumption if it is not possible to produce a force field that can be assumed to be transferable to general use.

For this situation to be improved it would require a greater depth of molecular properties, this is the case with the structure set investigated in chapter 8 and a further discussion of these points will be given at the end of that chapter.

One of the sections of the reduced final force field in which it is possible to have some confidence is the method of calculating the strain free angle values as this appears to produce realistic values for all combinations of atoms with only just over twice the number of parameters needed for the 'central only' and the 'central times substitution factor' options which have been shown to give considerably worse results. The small number of parameters involved also means that they can be optimised computationally, a situation that is impossible for the large number of parameters that would be required if a more full option was to be used for anything more than a few atom types.

Overall the conclusion is that the following parameter types can and cannot be reasonable optimised by the technique used in OPTPARAM for such structure sets investigated in this study:

Optimisable	Not Optimisable
Bond Length	Bond Stretch Force Constants
Strain Free Angle	Angle Bend Force Constants
Torsional Barriers	Non-bonded Interactions
Out-of-Plane Bending	

For those that did not optimise well for the current structure set the bond stretch force constants could be dramatically improved if infra-red frequency data was included or could simply be fixed to those values produced by the parameters taken from CHEMMIN. It is reasonable to think that the options will become more optimisable if a greater depth of experimental data is available. This is the situation that considered in the following chapter.

As mentioned above, the main movement in the hydrogen scaling factors for the angle bend parameters towards unreasonable levels took place in the final optimisation and this appears to have had a minimal effect on the residual factors. Because of this the main options for this parameter type are either to fix the parameters to the same values as in CHEMMIN, or to fix only the scaling factors to their CHEMMIN values and allow the other parameters to optimise.

Of the parameters that do not optimise to reasonable values using OPTPARAM the ones which cause the most difficulty are the non-bonded interactions as for many areas where molecular mechanics is used the need for accurate non-bonded interactions is essential. The possible approachs to overcome this problem are as follows:

1) Use values from other force fields, e.g. COSMIC, and then optimise the other parameters while keeping these values constant.
2) Use standard values for van der Wall Radii but allow the energy parameters to vary.
3) Optimise all non-bonded parameters but with a structure set where close non-bonded distances are available in significant numbers to reduce the production of atrefacts.

It is the third approach above that is effectively used in the study in the next chapter and the discussion of the suitability of OPTPARAM for optimising non-bonded parameters will continue there.

On the positive side it can be seen that there are several options which even for the deficient structure set used in this investigation show promise of helping to produce a general force field which does not require an excessive number of parameters. Of these the option that shows most promise and is new to this study is that of the calculation of strain free valence angles from a total angle.

7.7 References

1 J.G.Vinter, A. Davis and M.R.Saunders, *J. Comput.-Aided Mol. Design*, 1987, 1, 31.

2 D.N.J.White, J.N.Ruddock and P.R.Edgington, *Molecular Simulation*, 1989, **3**, 71.

3 Proteins: A Theoretical Perspective of Dynamics, Structure and Thermochemistry, C.L.Brooks, M.Karplus and B.Montgomery Pettitt, Advances in Chemical Physics, Volume LXXI. Page 34.

4 E.J.M. van Schaick, H.J.Geise, F.C.Mijlhoff and G.Renes, *J. Mol. Struc.*, 1973, **16**, 23.

5 P.Huber-Walchli and Hs.H.Gunthard, *Chem. Phys. Lett.*, 1975, **30**, 347.

6 Thermochemical Data of Organic Compounds, J.B.Pedley, R.D.Naylor and S.P.Kirby, Chapman and Hall Ltd., 1986, ISBN 0-412-27100-1.

7 A.L.Verma, U.F.Murphy and H.J.J.Bernstein, *J. Chem. Phys.*, 1974, **60**, 1540.

8 A.Almeninen, I.M.Anfinsen and A.Haaland, *Acta Chem. Scand.*, 1970, **24**, 43.

9 A.Almeninen, I.M.Anfinsen and A.Haaland, *Acta Chem. Scand.*, 1970, **24**, 43.
10 E.N.Marvell and R.S.Kitson, *J. Org. Chem.*, 1970, **35**, 338.
11 I.Hargittai, *Chem. Comm.*, 1971, 1499.
12 A.Yokozeki and K.Kochitsu, *Bull. Chem. Soc. Japan*, 1971, **44**, 1783.
13 S.Kondo, E.Hiroto and Y.Morino, *J. Mol. Spectroscopy*, 1968, **28**, 471.
14 A.Almeninen, G.G.Jacobson and H.M.Seip, *Acta Chem. Scand.*, 1969, **23**, 1495.
15 G.Dalligna and L.H. Toneman, *J. Mol. Struc.*, 1969, **1**, 177.
16 J.F.Chang and S.H.Bauer, *J. Amer. Chem. Soc.*, 1969, **91**, 1898.
17 J.Laane and R.H.Lord, *J. Chem. Phys.*, 1967, **47**, 4941.
18 I.Tokue, T.Fukuyama and K.Kuchitsu, *J. Mol. Struc.*, 1974, **23**, 33.
19 K.Kuchitsu, *J. Chem. Phys.*, 1966, **44**, 906.
20 L.H.Scharpen and V.W.Laurie, *J. Chem. Phys.*, 1963, **39**, 1732.
21 G.Gundersen, P.Murrey-Rust, D.W.H.Rankin, R.Seip and C.I.F.Watt, *Acta Chem. Scand.*, Ser. A, **A37(10)**, 823.
22 T.Shimanouchi, Y.Abe and K.Kuchitsu, *J. Mol. Struc.*, 1968, **2**,82.
23 A.Yokozeki and K.Kochitsu, *Bull. Chem. Soc. Japan.*, 1971, **44**, 1783.
24 B.A.Arbuzov and V.A.Naumov, *Doklady Akad. Nauk S.S.R.*, 1964, **154**, 376.
25 Molecular Structure By Diffraction Methods., Ed. G.A.Sim and L.E.Hutton (Specialist periodical reports), The Chemical Society, London, 1975, Vol 2, Page 8.
26 A.Immirzi and G.Allegra, *Atti Accad. Lincei. Rend. Classi Sci. fis. mat. nat.*, 1967, **43**, 338.
27 H.J.Geise, H.R.Buys and F.C.Mijlhoff, *J. Mol. Struc.*, 1971, **9**, 447.
28 M.Davis and O.Hassel, *Acta Chem. Scand.*, 1970, **17**, 1181.
29 M.Nakata and K.Kuchitsu, *J. Mol. Struc.*, 1982, **95**, 205.
30 S.H.Schei and R.Stolevik, *J. Mol. Struc.*, 1985, **128**, 171.
31 K.Hagen and K.hedberg, *J. Amer. Chem. Soc.*, 1973, **95**, 8266.
32 Q.Shen, *J. Mol. Struc.*, 1982, **97**, 133.
33 T.Nordtomme and K.Hagen, *J. Mol. Struc.*, 1985, **128**, 127.
34 *J.Mol.Struc.*, 1985, **128**, 151.
35 C.Kato, S.Konaka, T.Iijima and M.Kimura, *Bull. Chem. Soc. Japan*, 1969, **42**, 2148
36 T.Iijima, *Bull. Chem. Soc. Japan*, 1972, **45**, 3526.
37 M.Treatteberg, *Acta. Chem. Scand.*, 1974, **24**, 373.
38 D.E.Brown and B.Beagley, *J. Mol. Struc.*, 1977, **38**, 167.
39 J.R.Durig, S.E.Godbey and J.F.Sullivan, *J. Chem. Phys.*, 1984, **80**, 5893.

40 R.J.Abraham and G.Gatti, *J. Chem. Soc.* B, 1969, 1969.
41 D.Freisen and K.Hedburg, *J. Am. Chem. Soc.*, 1980, **102**, 3982.
42 K.Hagen, *J. Mol. Struc.*, 1985, **128**, 139.
43 S.Konaka, K.Kiyoto, K.Yamamoto, N.Suzuki and M.Kimura, *J. Mol. Struc.*, 1984, **125**, 143.
44 S.Yamamoto, M.Nakata and K.Kuchitsu, *J. Mol. Struc.*, 1985, **112**, 173.
45 M.Nakata, K.Kohata, T.Fukuyama and K.Kuchitsu, *J. Mol. Struc.*, 1980, **68**, 271.
46 M.Nakata, T.Fukuyama, C.J.Wilkins and K.Kuchitsu, *J. Mol. Struc.*, 1981, **71**, 195.
47 S.H.Schei and K.Hagen, *J. Mol. Struc.*, 1984, **116**, 249.
48 J.Dillon and H.J.Geise, *J. Mol. Struc.*, 1981, **72**, 247.
49 J.Dillon and H.J.Geise, *J. Mol. Struc.*, 1980, **69**, 137.
50 K.Tamagawa, R.L.Hilderbrandt and Q.Shen, *J. Am. Chem. Soc.*, 1987, **109**, 1380.
51 O.I.Sovik, O.Trongmo, K.Hagen, S.H.Schei, R.Stolevik and Q.Shen, *J. Mol. Struc.*, 1984, **118**, 1.
52 J.P.Guillory and L.S.Bartell, *J. Chem. Phys.*, 1965, **43**, 654.
53 B.Beagley, M.O.Jones and P.Yavari, *J. Mol. Struc.*, 1981, **71**, 203.
54 P.Vannuffel, L.van den Enden, L.van Alsenoy and H.J.geise, *J. Mol. Struc.*, 1983, **116**, 99.
55 H.Schei, K.Hagen, M.Traettenberg and R.Seip, *J. Mol. Struc.*, 1980, **62**, 121.
56 A.Immirzi and G.Allegra, *Atti Accad. naz. Lincei, Rend. Classe Sci. fis. mat. nat.*, 1967, **43**, 338.
57 S.Konaka, M.Takeuchi and M.Kimura, *J. Mol. Struc.*, 1985, **131**, 317.
58 F.C.Mijlhoff, G.H.Renes, K.Kohata, K.Oyanagi and K.Kuchitsu, *J. Mol. Struc.*, 1977, **39**, 241.
59 G.N.D.Al-Aldah, B.Beagley and M.O.Jones, *J. Mol. Struc.*, 1980, **65**, 271.
60 B.Beagley, M.O.Jones and M.A..Zanjanchi, *J. Mol. Struc.*, 1979, **56**, 215.
61 S.H.Schei and R.Seip, *Acta. Chem. Scand.*, 1984, **A38**, 345.
62 R.L.Hilderbrandt and S.H.Schei, *J. Mol. Struc.*, 1984, **118**, 11.
63 S.H.Schei and Q.Shen, *J. Mol. Struc.*, 1982, **81**, 269.
64 S.H.Schei, *J. Mol. Struc.*, 1983, **102**, 305.
65 S.Tsuchiya, *J. Mol. Struc.*, 1974, **22**, 77.
66 L.Fernholt and K.Kveseth. *Acta. Chem. Scand.*, 1978, **A32**, 63.
67 K.Kveseth, *Acta. Chem. Scand.*, 1975, **A29**, 307.
68 K.Hagen, K.Hedburg, J.Neisess and G.Gundersen, *J. Am. Chem. Soc.*, 1985, **107**, 341.

69 O.Trongmo, Q.Shen, K.Hagen and R.Seip, *J. Mol. Struc.*, 1981, **71**, 185.
70 O.I.Sovik, H.Schei, R.Stolevik, K.Hagen and Q.Shen. *J. Mol. Struc.*, 1984, **116**, 239.
71 Q.Shen, *J. Mol. Struc.*, 1984, **112**, 169.
72 O.Steinnes, Q.Shen and K.Hagen, *J. Mol. Struc.*, 1980, **66**, 181.
73 E.J.Jacob, *J. Mol. Struc.*, 1979, **52**, 63.
74 O.Steinnes, Q.Shen and K.Hagen, *J. Mol. Struc.*, 1980, **64**, 217.
75 S.Dingeseth, H.Schei and K.Hagen, *J. Mol. Struc.*, 1983, **102**, 45.
76 G.Gunderson, H.G.Thomassen, J.E.Boggs and M.L.Collins. *J. Mol. Struc.*, 1989, **213**, 1.
77 A.Spelbod, P.A.G.Huisman, F.C.Mijlhoff and G.H.Renes. *J. Mol. Struc.*, 1978, **44**, 159.
78 Q.Shen, R.L.Hilderbrandt and K.Hagen, J. Mol. Struc., 1980, **71**, 161.
79 P.A.G.Huisman, K.J.Klebe, F.C.Mijlhoff and G.H.Renes. *J. Mol. Struc.*, 1979, **57**, 71.
80 R.L.Hilderbrandt, A.L.Andreassen and S.H.Bauer, *J. Phys. Chem.*, 1970, **24**, 1586.
81 A.H.Lowrey, C.George, P.D'Antonio and J.Karle, *J. Mol. Struc.*, 1979, **53**, 189.
82 M.Traetteberg, G.Paulen, S.J.Syvin, Y.N.Panchenko, and V.I. Mochalov, *J. Mol. Struc.*, 1984, **116**, 141.
83 B.Beagley, R.K.Knowles and R.G.Pritchard, *J. Mol. Struc.*, 1987, **161**, 105.
84 S.Dingeseth, S.H.Schei and K.Hagen, *J. Mol. Struc.*, 1984, **116**, 257.
85 T.Sakurai, M.ishiyama, H.Takeuchi, K.Takeshita, K.Fukushi and S.Konaka, *J. Mol. Struc.*, 1989, **213**, 245.
86 J.H.M.Brake, R.A.J.Dreissen, F.C.Mijlhoff, G.H.Renes and A.H.Lowrey, *J. Mol. Struc.*, 1982, **81**, 277.
87 V.Mom, P.A.G.Huisman, F.C.Mijlhoff, and G.H.Renes, *J. Mol. Struc.*, 1980, **62**, 95.
88 P.A.G.Huisman and F.C.Mijlhoff, *J. Mol. Struc.*, 1979, **57**, 83.
89 P.A.G.Huisman and F.C.Mijlhoff, *J. Mol. Struc.*, 1979, **54**, 145.
90 P.A.G.Huisman, F.C.Mijlhoff and G.H.Renes, *J. Mol. Struc.*, 1979, **51**, 191.
91 N.L.Allinger, Y.H.Yuh and J.H.Lii, *J. Am. Chem. Soc.*, 1989, **111**, 8551.
92 C.R.C. Handbook of Chemistry and Physics, 1990, **9**, 95.

CHAPTER 8

Application of OPTPARAM To an X-Ray and Neutron Structure Set

8.1 Introduction

It was shown in chapter 7 that the computer optimisation of a molecular mechanics force field using OPTPARAM was susceptible to the insufficiencies of the structure set. A significant part of the problem is the use, exclusively, of structures determined by electron diffraction. Explicitly this difficulty arises from the fact that the papers from such studies give only a selection of the internal co-ordinates. To produce a reasonable force field for a greater number of atom types it was necessary to abandon the exclusive use of electron diffraction and, in addition, use structures determined by x-ray and neutron diffraction. The advantage in using these structures is that the Cartesian co-ordinates are available and so it is possible to generate all internal co-ordinates as well as many close non-bonded distances. The result of this is that it is possible to obtain a much larger number of structural properties than parameters to be optimised, which will reduce the likelihood of artefacts being produced in the optimised parameters.

Because of the large increase in the number of structures made available by this decision and the large amount of data available for each structure it was decided to increase the number of atom types used in the structures. This will also help in evaluating if the reduced parameter options result in a valid force field.

A large number of structures were obtained using the CSSR version of the Cambridge Structural Database[1]. A selection was taken and their co-ordinates obtained. Structures determined by neutron diffraction were given preference as these have more accurate hydrogen positions. Otherwise the selection was on a low R factor and for structures with as many different atom types as possible.

Unfortunately, the information obtained from the CSSR version of the database did not contain information relating to the published standard deviation of the atomic co-ordinates or the types of the atoms beyond their basic atomic species.

Due to time constraints when this data was being processed it was simply assumed that all non-hydrogen's had a standard deviation of 0.004Å and all hydrogen's had a standard deviation of 0.020Å. Using these variations the standard deviations of all the standard structural properties were calculated. As well as the calculation of bond lengths, valence angles and torsion angles, Non-bonded distances were included when the separation of the atoms involved was within 1.5 times the sum of their approximate van der Waal radii.

To simplify the input of data a program was produced using the modelling program CHEMMOD as a basis. This program read in the atom co-ordinates and connectivity table for a structure obtained from the CSSR database. The program used assumptions about bond lengths, angles and connectivities to estimate the

actual type of each atom. The user was then able to interactively change the types of the small number of atoms whose atom types had been wrongly set. After correcting any wrongly set atom types the user was prompted for any heat of formation information before the file was saved in a form that could be read directly into OPTPARAM. Finally before any files could be included in the structure set the structures were run through CHARGE2[2] to calculate the point charges. Some of the structures had non-zero charges and in these cases the structure was fed through the MNDO[3] program to produce the overall charge distribution as CHARGE2 cannot calculate charge distributions for ions.

The final structure set used in the optimisation during this investigation contained three different types of structures:

a) Electron diffraction structures from the previous investigation, mainly those for which some form of energy data was available. Because of the problem of bonds involving hydrogen's, as stated in section 2.3.4.2, all structural properties for bonds involving hydrogen used in the calculation of the structural residual factors were removed. These structures include not only structures from the earlier investigations but also other structures from the original scan of journals that did not fit into the categories covered in the earlier investigation.

b) X-ray and Neutron diffraction structures, produced as described above.

c) Heat of formation structures. These are structures for which essential heat of formation data was available but for which no structural data could be found. These were built using the CHEMMOD modelling package before being processed in the same way as the other structures.

The structure files produced by these three methods were then used in optimisations. Initially, the heat of formation data had to be excluded from the optimisation due to a small coding problem with the least squares procedure, but this was corrected well before the final optimisations took place.

Unfortunately, many of the new entries were stored in a directory that was not as regularly backed up as the main OPTPARAM directory structure and most of this data was lost when then the hard disk crashed at the industrial sponsor's site, where this work was taking place. Therefore it is only possible to draw limited conclusions from the optimisations.

8.2 Atom Types

The following is a list of the atom types used in this investigation.

H

C_{sp} C_{sp2} C_{arom} C_{sp3}

N_{sp} N_{sp2} N_{arom} N_{amide} N_{nitro} N_{sp3} N_{cation}

O_{sp2} O_{sp3} O_{anion}

F

P_{iii} P_{v}

S_{-ii-} $S_{ii=}$ S_{iv} S_{vi}

Cl

Br

I

Some of these types can refer to different bonding schemes. The following list gives the bonding schemes on which the optimisation took place:

C_{sp}	has one single and one triple bond.
C_{sp2}	has two single and one double bond.
C_{arom}	has one single and two aromatic bonds.
C_{sp3}	has four single bonds.
N_{sp}	has one triple bond.
N_{sp2}	has one single and one double bond.
N_{arom}	has two aromatic bonds.
N_{amide}	planar with three single bonds, one connected to a double bonded atom type.
N_{nitro}	has one single bond and two double bonds to oxygen's.
N_{sp3}	has three single bonds and is pyramidal.
N_{cation}	has four single bonds.
O_{sp2}	has one double bond.
O_{sp3}	has two single bonds.
O_{anion}	has one single bond.
P_{iii}	has three single bonds.
P_{v}	has one double and three single bonds.
S_{-ii-}	has two single bonds.
$S_{ii=}$	has one single bond.
S_{iv}	has one double and two single bonds.
S_{vi}	has two double and two single bonds.

8.3 Structure Set

The following is a list of the structures used in this limited optimisation. The first section contains those structures where the data was taken from electron diffraction studies. This section is divided into those structures for which an energy difference between isomer value is used in the optimisation and those used only for their structure and heat of formation residual factors. The second section relates to those structures taken from the CSSR database and section three contains the structures constructed for their heat of formation data only. Some structures in the second section contained more than one suitable structure, in these cases the structures were processed into separate files. In the list where a reference code is repeated the structure used in each case is underlined.

In the first two sections an asterisk at the end of the line indicates that that structure also contributes to the heat of formation residual factors. In the third section all structures contribute only to this residual factor. All heat of formation data was taken from the same source[4].

Appendix I contains the chemical diagrams and references for the X-ray and neutron diffraction structures given in the second section.

Section 1 Electron Diffraction Data

Part A Structures involved in Energy Difference Calculations

Compound Name

ANTI 1 BROMO PROPANE[5]	No. of Props. 0
GAUCHE 1 BROMO PROPANE	No. of Props. 0
ANTI 1 CHLORO 1 BROMO ETHANE[6]	No. of Props. 0
GAUCHE 1 CHLORO 1 BROMO ETHANE	No. of Props. 0
ANTI 1,1,2,2-TETRAFLUORO ETHANE[7]	No. of Props. 3
GAUCHE 1,1,2,2-TETRAFLUORO ETHANE	No. of Props. 1
ANTI 1,2-DIBROMO ETHANE[8]	No. of Props. 2
GAUCHE 1,2-DIBROMO ETHANE	No. of Props. 1
ANTI 1,2-DICHLORO ETHANE[9]	No. of Props. 2
GAUCHE 1,2-DICHLORO ETHANE	No. of Props. 1
ANTI 1,2-DIFLUORO ETHANE[10]	No. of Props. 3
GAUCHE-1,2 DIFLUORO ETHANE	No. of Props. 0
ANTI 2,3-DICHLORO-1,3-BUTADIENE[11]	No. of Props. 3
GAUCHE 2,3-DICHLORO-1,3-BUTADIENE	No. of Props. 1
ANTI 2,3-DICHLORO-1-PROPENE[12]	No. of Props. 4
GAUCHE 2,3-DICHLORO-1-PROPENE	No. of Props. 1
ANTI 2-BROMO-3-CHLORO PROPENE[13]	No. of Props. 3
GAUCHE 2-BROMO-3-CHLORO-PROPENE	No. of Props. 1

ANTI 2-BROMO-3-CHLORO-2-METHYL PROPANE[14]	No. of Props. 2
GAUCHE 2-BROMO-3-CHLORO-2-METHYL PROPANE	No. of Props. 1
ANTI TRANS-2-BUTENOYL CHLORIDE[15]	No. of Props. 1
SYN TRANS-2-BUTENOYL CHLORIDE	No. of Props. 6
SYN-SYN FUMARYL CHLORIDE[16]	No. of Props. 0
ANTI-ANTI FUMARYL CHLORIDE	No. of Props. 3
ANTI-SYN FUMARYL CHLORIDE	No. of Props. 1
ANTI-BUTANE[17]	No. of Props. 0
GAUCHE-BUTANE	No. of Props. 0
BOAT-CHAIR BICYCLO[3.3.1] NONANE[18]	No. of Props. 0
TWIN-CHAIR BICYCLO[3.3.1] NONANE	No. of Props. 0
CHLORO ACETYL CHLORIDE (ANTI)[19]	No. of Props. 5
CHLORO ACETYL CHLORIDE (GAUCHE)	No. of Props. 1
CIS 1-CHLORO-1,3-BUTADIENE[20]	No. of Props. 3
TRANS 1-CHLORO-1,3-BUTADIENE	No. of Props. 3
CIS BUT 2 ENE[21]	No. of Props. 1*
TRANS BUT 2 ENE	No. of Props. 1

Part B Structures not involved in Energy Difference Calculations

1,4 DIOXANE[22]	No. of Props. 12*
1,5 HEXADIYNE[23]	No. of Props. 5
1,6 DIBROMO 1,5 HEXADIYNE[24]	No. of Props. 6
2 NITRO PROPANE[25]	No. of Props. 6*
2-AMINOPYRIMIDINE[26]	No. of Props. 9
3-BUTENE NITRILE[27]	No. of Props. 5*
4,4 SULFANDIYL-BIS-THIOPHENOL[28]	No. of Props. 6
6,6 DINITRO 2,2 DIPHENIC ACID[29]	No. of Props. 11
ACETYLACETONE[30]	No. of Props. 12*
BENZENE[31]	No. of Props. 1*
BIPHENYL[32]	No. of Props. 7*
BIPHENYL SULPHONE[32]	No. of Props. 2
BIS(DIMETHYL PHOSPINO) METHANE[33]	No. of Props. 1
BIS(DISFLUOROTHIOPHOROL) ETHER[34]	No. of Props. 9
BIS(DISFLUOROTHIOPHOROL) METHANE[34]	No. of Props. 9
BIS(TRIFLUOROMETHYL)SULFOXYDIFLUORIDE[35]	No. of Props. 9
CARBONYLCYANIDE[36]	No. of Props. 7
HEXAFLUORO DIMETHYL SULPHIDE[37]	No. of Props. 4
TRIFLUORO METHYL SULFER MONOCHLORIDE[37]	No. of Props. 5
TRIFLUORO METHYL SULFER MONOFLOURIDE[37]	No. of Props. 5
DI-t-BUTYL SULPHIDE[38]	No. of Props. 15*
DIFLUORO BROMO PHOSPHENE[39]	No. of Props. 7

DIFLUORO CHLORO PHOSPHENE[39]	No. of Props. 7
DIFLUORO PHOSPHINE[39]	No. of Props. 7
DIMETHYL ETHER[40]	No. of Props. 3*
DIMETHYL OXYMETHANE[41]	No. of Props. 9*
DIMETHYL PEROXIDE[42]	No. of Props. 6*
CYANO BENZENE[43]	No. of Props. 6
DIPHENYL SULPHONE[44]	No. of Props. 4
FLUORO AMIDE SULPHOXIDE[45]	No. of Props. 7
FORMALDEHYDE[46]	No. of Props. 1*
TRI FLUOROSULPHOXY METHANE[45]	No. of Props. 10
HEXAFLUOROAZOMETHANE[47]	No. of Props. 5
IMIDO DISULPHONYL CHLORIDE[48]	No. of Props. 8
M CHLORO NITRO BENZENE[49]	No. of Props. 7
M DINITRO BENZENE[49]	No. of Props. 7*
MONOBROMO BENZENE[50]	No. of Props. 6*
MONOFLUORO BENZENE[51]	No. of Props. 6*
N,N,O, TRIMETHYL HYDROXYAMINE[52]	No. of Props. 7
OXALIC ACID[53]	No. of Props. 10*
NAPHTHALENE[54]	No. of Props. 4*
P BROMO NITRO BENZENE[55]	No. of Props. 12
p-ANISALDEHYDE[56]	No. of Props. 11
PHOSPHORYL BROMIDE[57]	No. of Props. 4
SULPHURYL CHLORIDE[58]	No. of Props. 5
TETRAMETHYL HYDRAZINE[59]	No. of Props. 10
TETRAMETHYL UREA[60]	No. of Props. 8
THIO CARBONYL DIBROMIDE[61]	No. of Props. 3
THIO CARBONYL DIFLUORIDE[62]	No. of Props. 3
THIO PHENESULPHONYLCHLORIDE[63]	No. of Props. 12
THIO PHOSPHORYL BROMIDE[57]	No. of Props. 4
METHYL (VINYL)-N-NITRAMINE[64]	No. of Props. 13
QUINUCLIDINE[65]	No. of Props. 18*
THIOCARBONYL CHLORIDE[66]	No. of Props. 3
TRIFLUOROMETHYL HYPOCHLORATE[67]	No. of Props. 5
TRIOXACYCLO PENTANE[68]	No. of Props. 14
TRIS (TRI FLUORO METHYL) PHOSPHINE[69]	No. of Props. 9

Section 2 X-Ray and Neutron Data	Refcode	
ACETAMIDE (NEUTRON STUDY)	ACEMID03	No. of Props. 32
2,4,5,6(1H,3H)-PYRIMIDINETETRONE	ALOXAN11	No. of Props. 61
AMMONIUM HYDROGEN 1-MALATE	AMHMAM	No. of Props. 70
AMMONIUM HYDROGEN D-TARTRATE	AMHTAR01	No. of Props. 76

AMMONIUM HYDROGEN D-TARTRATE	AMHTAR01	No. of Props. 10
AMINOMALONIC ACID	AMMALA	No. of Props. 64
2-AMINOPHENOL	AMPHOM02	No. of Props. 75
P-THIOCYANO-ANILINE	ANLINT20	No. of Props. 79
L-ASPARAGINE MONOHYDRATE (NEUTRON STUDY)	ASPARM02	No. of Props. 88
5-ACETAMIDO-1,3,4-THIADIAZOLE-2-SULFONAMIDE	ATDZSA	No. of Props. 84
1,1'-AZO-BIS(CARBAMIDE)	AZDCAR01	No. of Props. 44
DITHIOMALONODIAMIDE	BASRES	No. of Props. 58
2-AMINO-1,3-THIAZOLE	BAWKEP10	No. of Props. 42
BETAINE HYDROCHLORIDE	BETANC01	No. of Props. 128
2-(TRICHLOROPHOSPHAZO)-PERCHLOROPROPANE	BEZBIR	No. of Props. 79
TETRATHIAFULVALENE P-DINITROBENZENE	BIRDIP	No. of Props. 85
TETRATHIAFULVALENE P-DINITROBENZENE	BIRDIP	No. of Props. 67
D,L-BETA-CARBOXYASPARTIC ACID MONOHYDRATE	BIXMOK	No. of Props. 113
DIBROMO-(MALEIC ACID)-THIOANHYDRIDE	BMLTAA	No. of Props. 41
DI-IODO-(MALEIC ACID)-THIOANHYDRIDE	BMLTAB	No. of Props. 41
1,1,4,4-TETRAFLUOROCYCLOHEXANE	BOMJIW	No. of Props. 123
1,1-DITHIO-OXALIC ACID-2-AMIDE-1-S-METHYL ESTER	BOZKUW	No. of Props. 47
THIO-OXALIC ACID-1-AMIDE-2-O-ETHYL ESTER	BOZLAD	No. of Props. 66
BROMOMALONAMIDE	BRMALA	No. of Props. 38
2,3,6,7-TETRABROMONAPHTHALENE	BRNPHL	No. of Props. 73
1,1,2,2-ETHANETETRACARBONITRILE (NEUTRON STUDY)	BUGKIX02	No. of Props. 42
DIMETHYLTHIOPHOSPHINIC ACID	BUHZEJ	No. of Props. 51
1,4-DIHYDRO-1,2,4,5-TETRAZINE	BUSFUQ	No. of Props. 43
S-METHYL DITHIOCARBAMATE HEMIHYDRATE	BUXCEC	No. of Props. 35
TETRAMETHYLAMMONIUM NITROSOLATO-BENZOATE	BZNTSL10	No. of Props. 106
BENZOXAZOLIN-2-ONE	BZOXZO	No. of Props. 80
(E)-3-HEXENEDINITRILE	CADVEI	No. of Props. 55
1,4.3,6-DIANHYDRO-D-GLUCITOL	CEKPIR	No. of Props. 136
CIS,CIS-2-METHYL-3-HYDROXY-4-IODOMETHYL -GAMMA-BUTYROLACTONE	CEPRUK	No. of Props. 119
N-METHYLAMINOMETHANESULFONIC ACID	CEXMOH	No. of Props. 69
IMIDAZOLINE-2-THIONE HEMIHYDRATE	CEZXEK	No. of Props. 20
(-)-CHLOROSUCCINIC ACID	CLSUCC	No. of Props. 67
2,2A,4,4A,6,6A-HEXAHYDRO-1,3,5-TRITHIA-7 -AZACYCLOPENTA(CD)PENTALENE	COGHIP	No. of Props. 135
O-METHYL-THIOCARBAMATE	CONFAM	No. of Props. 37
1,2,3-PROPANETRIOL TRINITRATE	CORYIR	No. of Props. 99
(2R,3S)-2-HYDROXYMETHYL-3- -HYDROXYPYROLINE HYDROCHLORIDE	CUZMEP	No. of Props. 133
S-CARBOXYMETHYL-L-CYSTEINE	CXMCYT	No. of Props. 106

2-CYANOGUANIDINE	CYAMPD03	No. of Props. 35
CYANURIC ACID (NEUTRON STUDY)	CYURAC12	No. of Props. 59
THIENO(3,4-F)-1,2,3,4,5-PENTATHIEPIN	DAHDAR	No. of Props. 58
L-ASPARAGINE L-ASPARTIC ACID MONOHYDRATE	DAYREA01	No. of Props. 91
3,6-DIETHYL-1,2,4,5-TETRATHIA-3,6 -DIAZACYCLOHEXANE	DAZVIJ	No. of Props. 44
2,6-DIMETHYL-4H-1-THIAPYRAN-4-ONE	DEFYES	No. of Props. 86
2-AMINOPROPENENITRILE	DESYOP	No. of Props. 29
DIGLYCINIUM SULFATE MONOHYDRATE	DGLSLM10	No. of Props. 10
DL-SERINE (NEUTRON STUDY)	DLSERN11	No. of Props. 71
N,N-DIMETHYLANILINE HEXAFLUOROBENZENE COMPLEX	DMAFBZ	No. of Props. 63
DIMETHYLGLYOXIME	DMEGLY01	No. of Props. 67
1,3-DIMETHYLIMIDAZOLE-2(3H)-THIONE	DMIMZT10	No. of Props. 81
TRANS-3,6-DIMETHYL-THIENO(3,2-B)THIOPHENE -2,5(3H,8H)-DIONE	DMTTDO	No. of Props. 122
ANILINIUM IODIDE	DOVVIT	No. of Props. 59
METHYLENE-BIS(PHOSPHINIC ACID)	DOXHED	No. of Props. 56
1-HYDROXYMETHYL-3,6-DIMETHYL-1,6-DIHYDRO -1,2,3,5-TETRAZINE	DUDMUK	No. of Props. 94
3-IODO-BENZAMIDE	DUMNUU	No. of Props. 42
1-METHYL-3-ETHYL-IMIDAZOLIUM IODIDE	DUVZAV	No. of Props. 92
ETHANOL	ETANOL	No. of Props. 35*
MONOFLUOROACETAMIDE (NEUTRON STUDY)	FACETA01	No. of Props. 33
MONOFLUOROACETIC ACID	FACETC10	No. of Props. 27
4-BROMO-8-OXA-6-AZABICYCLO(3.2.1)OCTAN-7-ONE	FANPAL	No. of Props. 126
2,5-DIAMINO-1,3,4-THIADIAZOLE	FELRUJ	No. of Props. 36
METHYLIMINO-DIACETIC ACID	FENTOH	No. of Props. 110
TRIS(DICHLOROMETHYL)AMINE	FESCEL	No. of Props. 66
2-ALLYLAMINO-CDELTA-2--THIAZOL-4-ONE	FIVPIJ	No. of Props. 90
S-DIFORMYLHYDRAZIDE	FOMHAZ13	No. of Props. 39
N-HYDROXY-METHANEIMIDAMIDE (NEUTRON STUDY)	FORAMO01	No. of Props. 24
FORMIC ACID	FORMAC01	No. of Props. 11
DITHIENO(3,4-B,3',4'-E)(1,4)DITHIINE -1,3,5,7-TETRAONE	FOWVES	No. of Props. 93
TRANS-BETA-2-FURYL-ACRYLAMIDE	FUACAM	No. of Props. 81
GLYOXIME (NEUTRON STUDY)	GLOXIM11	No. of Props. 31
GLYCINE (NEUTRON STUDY)	GLYCIN15	No. of Props. 42
GLYCYLGLYCINE PHOSPHATE MONOHYDRATE	GLYGLP	No. of Props. 23
GUANIDINIUM 4-AMINO-3,5,6 -TRICHLOROPICOLINATE	GUPICL10	No. of Props. 63
L-HYDANTOINO(C)-1,4-PERHYDROTHIAZINE	HYDTZA	No. of Props. 119
IMIDAZOLE (NEUTRON STUDY)	IMAZOL06	No. of Props. 38

IMINODIACETIC ACID HYDROBROMIDE	IMDACB	No. of Props. 92
SQUARIC ACID (NEUTRON STUDY)	KECYBU06	No. of Props. 46
L-ALANINE (NEUTRON STUDY)	LALNIN12	No. of Props. 65
L-GLUTAMIC ACID (NEUTRON STUDY)	LGLUAC03	No. of Props. 104
TRIMETHYLAMMONIUM 7,7,8,8-TETRACYANOQUINO DIMETHANE-TRI-I-IODIDE	MATCQI09	No. of Props. 95
N-METHYLBENZAMIDE	MBNZAM10	No. of Props. 102
MELAMINE	MELAMI01	No. of Props. 44
MESO-ETHANE-1,2-BIS(METHYL SULFOXIDE)	METMSX	No. of Props. 88
6-METHYLMERCAPTOPURINE TRIHYDRATE	MMCPUR	No. of Props. 84
N-ACETYL-L-CYSTEINE (NEUTRON STUDY)	NALCYS02	No. of Props. 96
Z-1-(5-NITRO-2-FURYL)-2-THIOCYANATO-ETHYLENE	NFMLEB	No. of Props. 78
NITROACETIC ACID	NTRACD	No. of Props. 41
2-NITROGUANIDINE (NEUTRON STUDY)	NTRGUA01	No. of Props. 43
1-OXO-4-METHYL-2,6,7-TRIOXA-1-PHOSPHABIC (2.2.1)HEPTANE	OMOPCP	No. of Props. 103
OXALIC ACID DIHYDRATE	OXACDH11	No. of Props. 27
PERCHLORO-ALL-CIS-TRICYCLO(5.2.1.0-4,0-) DECA-2,5,8-TRIENE	PCLDTR	No. of Props. 163
TETRAFLUORO-TEREPHTHALODINITRILE	PFTENT11	No. of Props. 65
P-PHENYLENEDIAMINE DIHYDROCHLORIDE	PHNDMC11	No. of Props. 93
O-PHENYLENEDIAMINE DIHYDROCHLORIDE	PHNDMO	No. of Props. 94
PENTACHLOROBENZENE	PNCLBZ	No. of Props. 63
PUTRESCINE DIPHOSPHATE	PUTRDP02	No. of Props. 20
2-PYRIDONE (NEUTRON STUDY)	PYRIDO04	No. of Props. 57
2-THIOPYRIDONE (NEUTRON STUDY)	PYRIDS02	No. of Props. 33
SEMICARBAZIDE HYDROCHLORIDE [NEUTRON STUDY]	SECAZC02	No. of Props. 35
L-O-SERINE PHOSPHATE	SERPOP01	No. of Props. 101
D-(+)-TARTARIC ACID	TARTAC05	No. of Props. 81
(+)-(2R,3R)-TARTARIC ACID	TARTAC23	No. of Props. 85
TETRACYANOETHYLENE (NEUTRON STUDY)	TCYETY02	No. of Props. 25
2',2",2'"-PHOSPHINETRIYL-TRIACETONITRILE	TCYMPH02	No. of Props. 60
TRIGLYCINE SULFATE	TGLYSU07	No. of Props. 43
THIOACETAMIDE (NEUTRON STUDY)	THACEM01	No. of Props. 34
THIOCYTOSINE	THCYTO10	No. of Props. 64
THIOUREA	THIOUR08	No. of Props. 28
5,6-DIHYDROTHIAZOLO(2,3-C)(1,2,4)THIADIAZOL-3-ONE	THZOLA	No. of Props. 73
TETRAMETHYLENE PHOSPHORIC ACID	TMEYPH	No. of Props. 101
THIOSEMICARBAZIDE HYDROCHLORIDE	TSEMHC	No. of Props. 48
TRITHIO-OXALIC ACID S,S'-DIMETHYL ESTER	TTOXDM	No. of Props. 52
CIS-2,5,7-TRIOXA-BICYCLO(4.3.0)NONAN-8-ONE	TXBNON	No. of Props. 115
UREA (NEUTRON STUDY)	UREAXX14	No. of Props. 28

ALPHA-L-XYLOPYRANOSE (NEUTRON STUDY)	XYLOSE02	No. of Props. 122
0.8-DEUTERO-BETA-L-ARABINOSE (NEUTRON STUDY)	YYYYAA	No. of Props. 124
L(+)-GLUTAMIC ACID HYDROIODIDE	ZZZBUV01	No. of Props. 107
M-IODOBENZOIC ACID	ZZZOAE01	No. of Props. 79
1-AMINO-1-METHYLETHYL-PHOSPHONIC ACID MONOHYDRATE	ZZZSPS01	No. of Props. 103
TRIMETHYLAMMONIOSULFONATE	ZZZVPQ01	No. of Props. 112

Section 3 Heat of Formation Data

DIETHYL AMINE
DIETHYL ETHER
ISOPROPYL METHYL ETHER
DIMETHYL AMINE
TRIMETHYL AMINE
ACETONITRILE
BROMONITRILE
ETHYLAMINE
NITRO ETHANE
ETHYL HYDROSULFIDE
METHYL AMINE
ACETIC ACID
NITRO METHANE
METHANOL
OXYNITROSO METHANE
METHYL SULFER CYANIDE
METHYL HYDROSULFIDE
OXYDICHLORIDE
CYANIDE
HYDROGEN SULPHIDE
METHYL FORMATE
IODONITRILE
HYDRAZINE
TRIFLUORO AMINE
AMMONIA
NITROSO CHLORIDE
PROPIONITRILE

Table 8.1 shows the number of parameters for each property type. These are compared with the number of similar properties for the investigation carried out in Chapter 7.

Table 8.1 A Comparison of Property and Structure Numbers Between Studies

Property Type	This Study	Chapter 7
Distance	2934	249
Valence Angle	2910	233
Torsion Angle	3328	51
Energy Difference	17	17
Heat of formation	46	71
Number of Structures	246	147

This table shows the advantages and disadvantages of using structures determined by methods other than electron diffraction. On the positive side it can be seen that the number of structural properties has gone up by 12 to 65 times the numbers in the Chapter 7 study. On a more negative side the number of energy difference properties has remained constant and the number of heat of formation properties has fallen. The reason for this is that it was only possible to find heat of formation data for a smaller percentage of the CSSR data. As a result of this it was necessary to include the extra structures listed above which contribute to the heat of formation data only. Even with these the number of heat of formation cases is still smaller than that in the Chapter 7 study. The reason why the number of energy difference cases stayed constant is that there is no standard reference work for these measures, as there was for the heat of formation data. Thus besides a few standard values these numbers were obtained for the electron diffraction studies. With data coming from the structural database no such information was readily available as it had been in the electron diffraction papers and with limited time it was not possible to attempt to find any such information for the structures under consideration.

8.4 Optimisation

In this investigation the same form of the force field was used as in the final optimisation for the second expanded structure set in Chapter 7. This configuration is as follows:

Bond Lengths:	Full
Bond Strength Force Constants:	Calculated
Strain Free Valence Angles:	Calculated from Total Angle
Angle Bend Force Constants:	Calculated
Torsional Barriers:	Central Two
Non-Bonded Parameters:	Cosmic[70]
Out of Plane Bending:	One Value per Atom Type

A difficulty that arises from using structures with so many structural properties is that the size of the COMPARE file becomes very large, in the region of 1.2 MBytes. It was because of this that the routine for writing the COMPARE file was adapted to only show those structural properties that fall outside their experimental error. Even with this option invoked the size of final COMPARE file is still in the region of 360Kbytes. The summary, heat of formation and energy difference data for the final optimisation are given later in this chapter as they are relatively concise. The rest of the COMPARE file is contained in appendix J that can be found on the microfiche and runs to 93 pages.

Because of the large amount of data and the fact that, unlike the final optimisation in chapter 7, the last optimisation did not have time to settle fully it is difficult to examine the result of the optimisation on individual structures, so the following discussions will generally refer only to the overall trends of the optimisation.

8.5 Results

8.5.1 Residual Factor Movement Trends During Optimisation

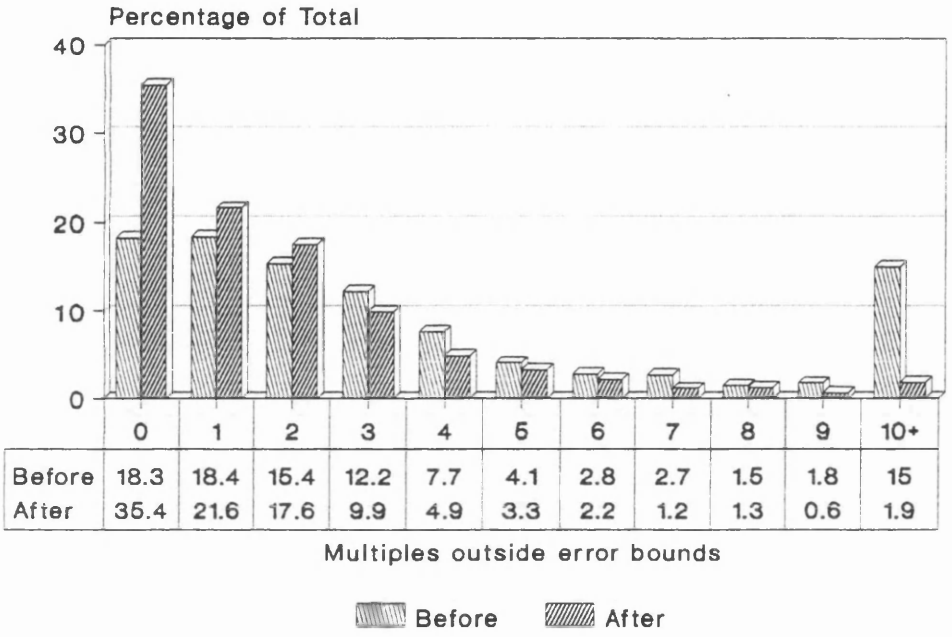
At the time when optimisation took place the data set was constantly being developed, i.e. the structure set was being altered, mainly by the addition of extra structures, as the investigation progressed. The list of structures above represents the final structure set for which the final results refer.

Because the structure set used throughout was not constant it is not possible to do an exact comparison between the initial force field and that which was produced after several weeks of, mainly overnight, optimisation. In an attempt to give an idea of the general effectiveness of the optimisation a broader measure has had to be used. What has been done in the following graphs is to examine the distribution of the various properties for the initial run where the information was available and the final run as a percentage of the total in each case. This then allows the production of distributions of the properties in these cases such that the distributions are comparable. The only exception to this is the Energy Difference graph where the same structures and differences used throughout.

Examining the distance graph, figure 8.1, the effect of the optimisation is very apparent. Before optimisation a significant number of bond lengths were more than 10 times the error out from the experimental value, after optimisation this number had dropped to almost an eighth of this number, 1.9% as opposed to 15%. For the final optimisation it would be necessary to include all the bands down to four times the error bounds before the same percentage could be reached.

Figures 8.1 & 8.2 Distance and Torsion Angle Residual Factor Distributions

Distances For X-Ray & Neutron Set

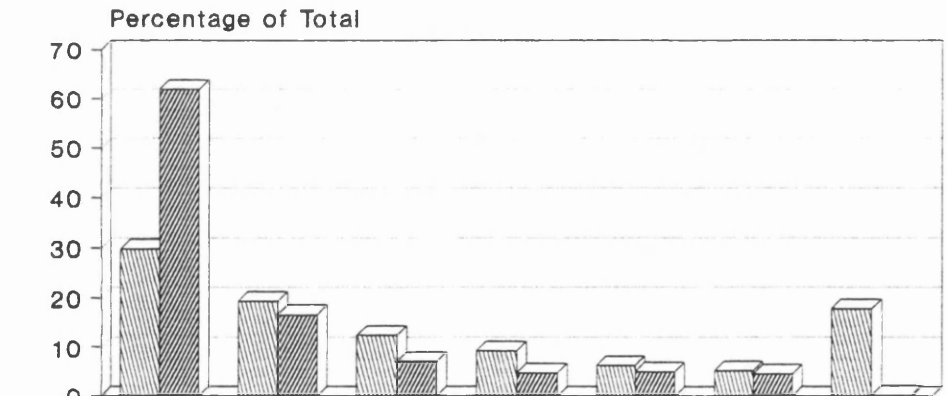


	0	1	2	3	4	5	6	7	8	9	10+
Before	18.3	18.4	15.4	12.2	7.7	4.1	2.8	2.7	1.5	1.8	15
After	35.4	21.6	17.6	9.9	4.9	3.3	2.2	1.2	1.3	0.6	1.9

Multiples outside error bounds

Before After

Torsion Angles For X-Ray & Neutron Set

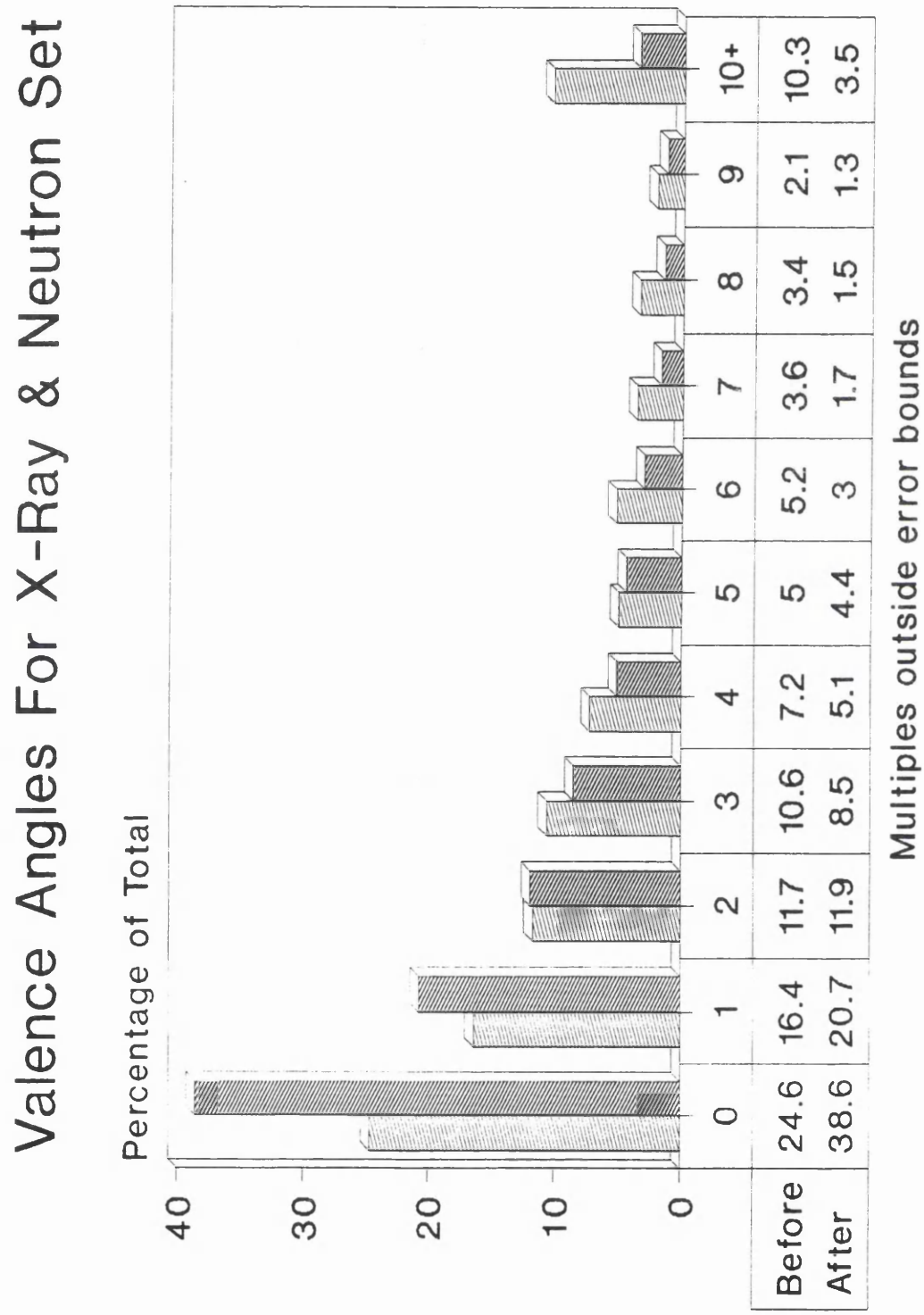


	0	1	2	3	4	5	6+
Before	29.8	19.3	12.4	9.3	6.3	5.2	17.7
After	62	16.6	7.1	4.7	5	4.5	0.2

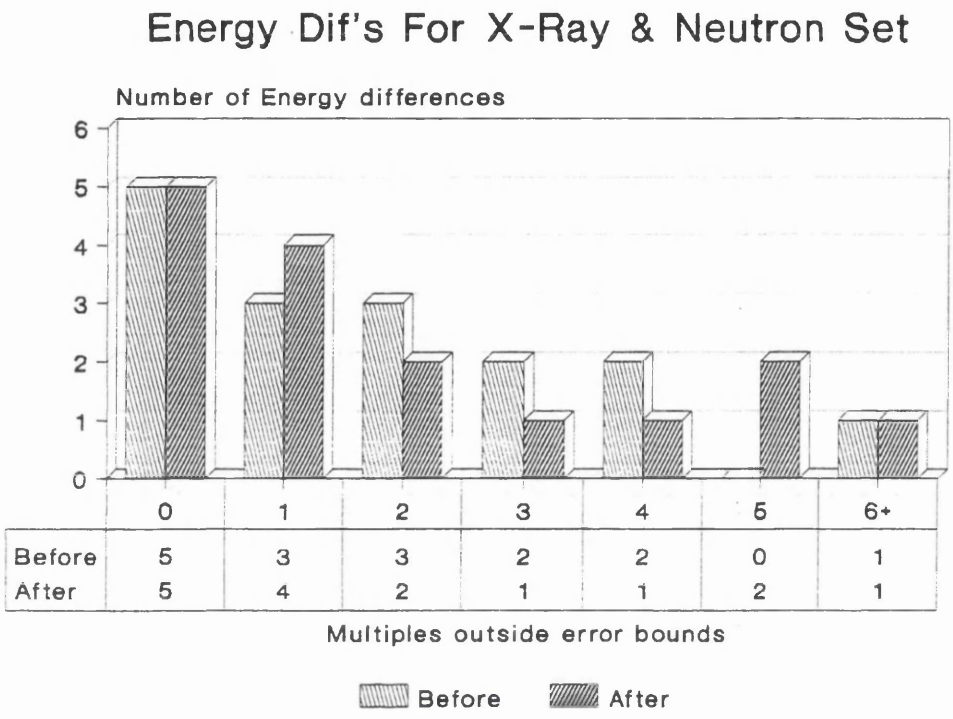
Multiples outside error bounds

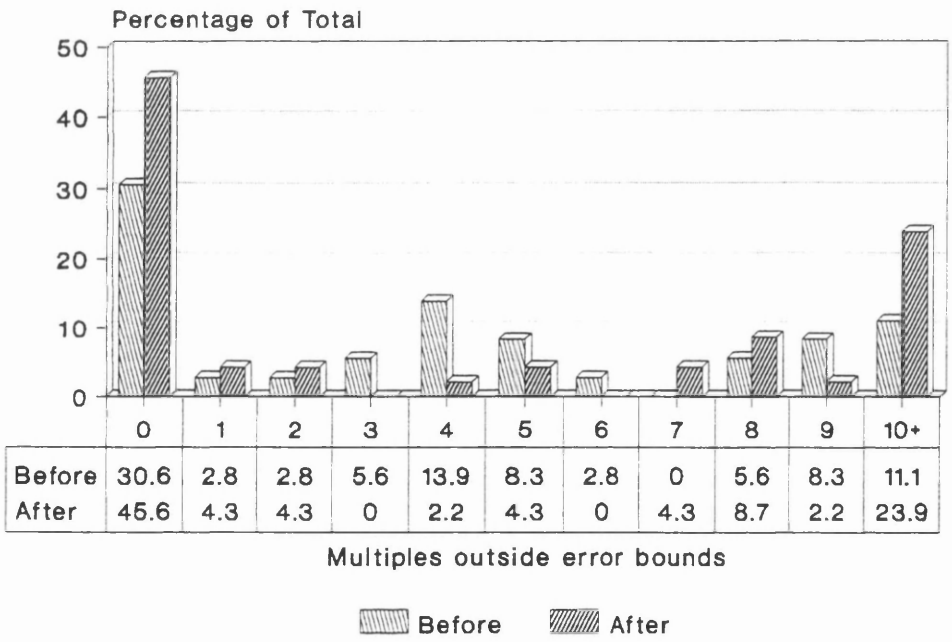
Before After

Figure 8.3 Valence Angle Residual Factor Distribution



Figures 8.4 & 8.5 Energy Residual Factor Distributions





This significant shift in the property distribution is also evident in figure 8.3, the valence angle graph. In this case the number of angles greater than 10 times the error has declined from 10.3% to 3.5%. It is also noticeable that all the error bands where the error was 3 times or greater show a net reduction in the number of angles, whereas all those smaller than this have a net increase. Of these, the greatest increase is in the number of valence angles that are within the experimental error, which for the final optimisation containing 12% more of the total number of angles than at the beginning.

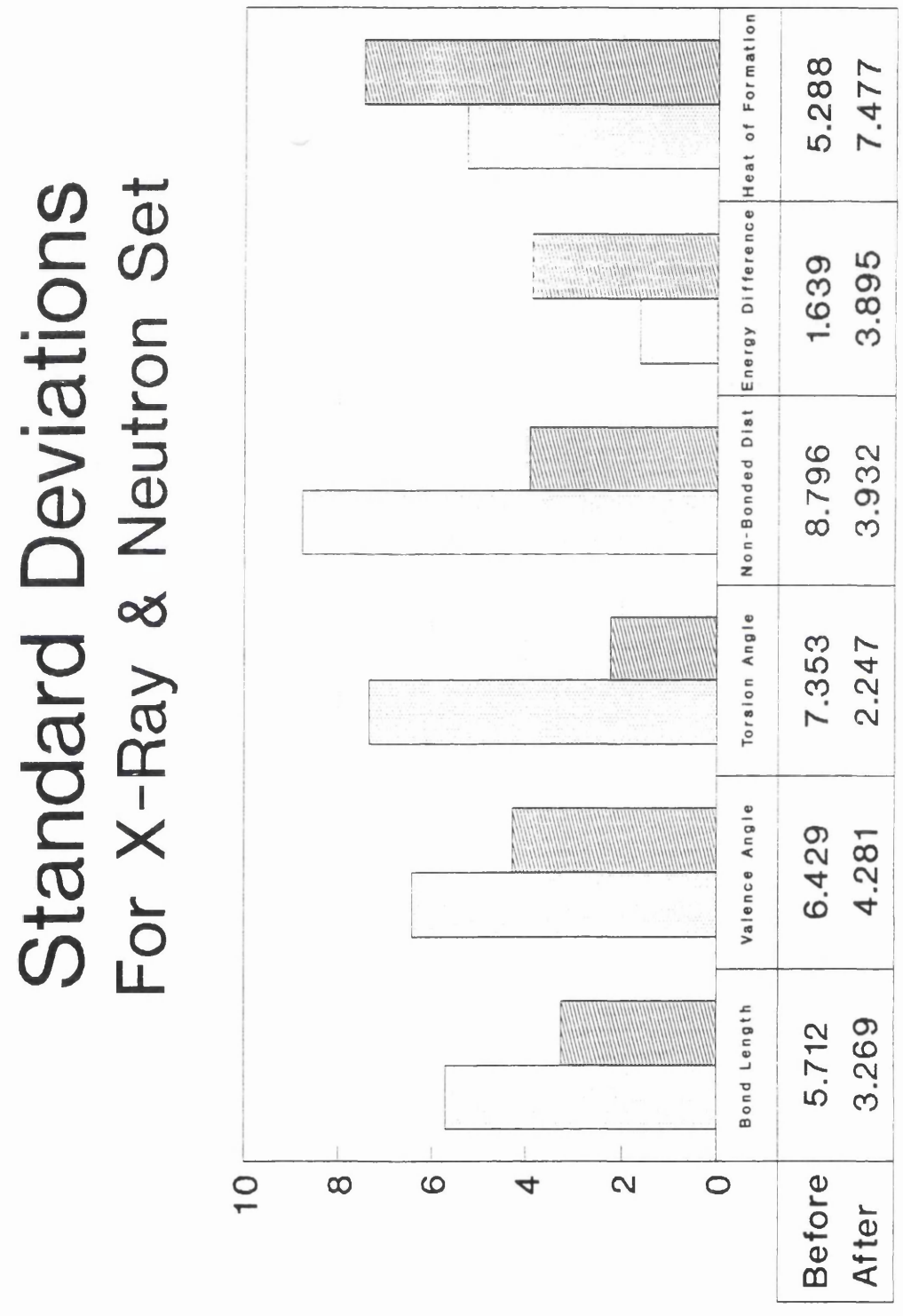
The movement of the results towards the low error bands is shown most dramatically in figure 8.2 for the torsion angles. This shows a 88 fold decrease in those entries with a difference to error ratio greater than 5. It must though, be remembered, that these runs were done with the additional energy component from those torsion's where the calculated angle was over 5 times the error out from the central experimental value. After taking this into consideration it is encouraging to find that all the error bands except the 'within error' band show lower values, whilst the 'within error' band has more than doubled in size from the initial optimisation. Thus the vast majority of the torsion angles will not have been affected by the additional energy component but still moved towards the 'within error' band.

The energy difference graph, figure 8.4, does not show any of this dramatic improvement. In fact, overall, the results are slightly worse. Because of the relative difficulty between obtaining this type of data compared to the structural and heat of formation data it should be noted that the sample size is now much smaller by comparison to these other methods. Thus there is little that can be reasonably deduced from these results.

The heat of formation data, figure 8.5, is a strange mixture where those residual factors that started at a high error moved outwards to become even worse whilst those in the middle move towards the lower end, resulting in 15% more residual factors being present in the 'within error' band at the end of the final optimisation. This unusual behaviour is almost certainly due to the problems arising from the use of the geometric enthalpy increments. It was shown in the previous chapter that the worse results were from those structures that used few geometric enthalpy increments.

All these trends can be seen in Figure 8.6 which shows the standard deviations for the 'Before' and 'After' cases discussed above. Note that by examining the magnitude of the distances under consideration it has been possible to extract from the results file those distances that will almost certainly correspond to

Figure 8.6 Standard Deviations of Residual Factors for X-Ray and Neutron Set



the non-bonded close contacts and the results for these are shown here in a separate column.

Figure 8.6 shows that the standard deviation for all the structural residual factors has come down over the course of the optimisation, whilst the energy residual factors show similar sized rises.

This implies that, unlike in Chapter 7, the weighting may have been too much in favour of the structural residual factors leaving the energy residual factors to become more unrealistic as the optimisation progressed. From the structural perception it is encouraging that, for the bond length and non-bonded interactions, the separation in the standard deviations for these factors was reduced from 3.084 to 0.663.

8.5.2 Final Results

The following is an extract from the final COMPARE file (appendix J). It lists the overall properties as well as the results for the heats of formation and energy differences. Comments are given in square '[]' brackets.

[The following averages relate to the actual properties, thus, for example, the average difference between the calculated valence angles and the central experimental value is 2.4087° compared to an average experimental error of 2.0963°]

```
AVERAGE BOND LENGTH DIFFERENCE          :-   0.0581    AVERAGE ERROR  :-   0.0380

AVERAGE BOND ANGLE DIFFERENCE           :-   2.4087    AVERAGE ERROR  :-   2.0963

AVERAGE TORSION ANGLE DIFFERENCE        :-   5.3090    AVERAGE ERROR  :-   5.5671

AVERAGE CONFORMER DIFFERENCE            :-   0.5292    AVERAGE ERROR  :-   0.3084

AVERAGE HEAT OF FORMATION DIFFERENCE    :-   1.2911    AVERAGE ERROR  :-   0.2499
```

[The following are the final residual factors that are combined to give to total, also show below. Note again how much larger the structural difference residual factor is when compared to the other two factors]

```
STRUCTURE DIFFERENCE :-   6.0417

CONFORMER ENERGY     :-   0.2355

HEAT OF FORMATION    :-   0.3260


TOTAL                :-  35.2443
```

[The following is the Heat of Formation data. Note that as with the similar table in chapter 7 that this data is taken directly from the COMPARE file and so the trailing zero's should be ignored when considering the number of significant figures for the experimental data]

```
CIS BUT 2 ENE                                *
CALC :    -1.1834  EXPT :    -1.6970  ERROR :     0.2400  DIF :     0.2736
ETHANOL (AT 87 DEG.K)
CALC :   -51.1168  EXPT :   -56.1900  ERROR :     0.1000  DIF :     4.9732
CYANIDE
CALC :    32.3000  EXPT :    32.3000  ERROR :     0.2000  DIF :     0.0000
ACETIC ACID
CALC : -103.8236  EXPT : -103.3100  ERROR :     0.1000  DIF :     0.4136
ETHYL HYDROSULFIDE
CALC :    -9.9982  EXPT :   -10.9500  ERROR :     0.1000  DIF :     0.8518
ETHYLAMINE
CALC :   -13.1357  EXPT :   -11.2700  ERROR :     0.1000  DIF :     1.7657
DIMETHYL AMINE
CALC :    -3.5761  EXPT :    -4.4100  ERROR :     0.1000  DIF :     0.7339
NITRO ETHANE
CALC :   -22.4657  EXPT :   -23.5600  ERROR :     0.1000  DIF :     0.9943
METHYL SULFER CYANIDE
CALC :    38.3000  EXPT :    38.3000  ERROR :     0.2000  DIF :     0.0000
TRIMETHYL AMINE
CALC :    -5.8808  EXPT :    -5.8100  ERROR :     0.1000  DIF :     0.0000
ISOPROYL METHYL ETHER
CALC :   -58.8370  EXPT :   -60.2600  ERROR :     0.1000  DIF :     1.3230
METHYL AMINE
CALC :    -5.0073  EXPT :    -5.4900  ERROR :     0.1000  DIF :     0.3827
DIETHYL ETHER
CALC :   -15.8093  EXPT :   -19.8600  ERROR :     0.1000  DIF :     3.9507
DIETHYL AMINE
CALC :   -17.8216  EXPT :   -17.0700  ERROR :     0.1000  DIF :     0.6516
OXYDICHLORIDE
CALC :    19.2000  EXPT :    19.2000  ERROR :     0.2000  DIF :     0.0000
AMMONIA
CALC :   -10.0784  EXPT :   -10.9730  ERROR :     0.0840  DIF :     0.8106
NYDRAZINE
CALC :    22.8000  EXPT :    22.8000  ERROR :     0.2000  DIF :     0.0000
TRIFLUORO AMINE
CALC :   -29.8000  EXPT :   -29.8000  ERROR :     0.2000  DIF :     0.0000
NITROSO CHLORIDE
CALC :    12.3600  EXPT :    12.3600  ERROR :     0.1000  DIF :     0.0000
HYDROGEN SULFIDE
CALC :    -5.8238  EXPT :    -4.9300  ERROR :     0.1000  DIF :     0.7938
NITROMETHANE
CALC :   -22.0433  EXPT :   -17.8600  ERROR :     0.1000  DIF :     4.0833
OXYNITROSO METHANE
CALC :   -16.5000  EXPT :   -16.5000  ERROR :     0.2000  DIF :     0.0000
BROMONITRILE
CALC :    44.5000  EXPT :    44.5000  ERROR :     0.2000  DIF :     0.0000
IODONITRILE
CALC :    53.9000  EXPT :    53.9000  ERROR :     0.2000  DIF :     0.0000
```

```
METHANOL
CALC :   -43.2796  EXPT :  -38.7200  ERROR :    0.1000  DIF :    4.4596
METHYL HYDROSULFIDE
CALC :    -4.5041  EXPT :   -5.3400  ERROR :    0.1000  DIF :    0.7359
METHYL FORMATE
CALC :   -83.1864  EXPT :  -83.7000  ERROR :    0.2000  DIF :    0.3136
DI-t-BUTYL SULPHIDE
CALC :   -44.9276  EXPT :  -39.9830  ERROR :    0.3820  DIF :    4.5626
ACETYLACETONE
CALC :   -90.9050  EXPT :  -90.9050  ERROR :    0.4300  DIF :    0.0000
3-BUTENE NITRILE
CALC :    37.1764  EXPT :   37.6900  ERROR :    0.2870  DIF :    0.2266
QUINUCLIDINE
CALC :    -0.5261  EXPT :   -1.0030  ERROR :    0.3100  DIF :    0.1669
2 NITRO PROPANE
CALC :   -30.1110  EXPT :  -33.2000  ERROR :    0.2150  DIF :    2.8740
DIMETHYL PEROXIDE
CALC :   -30.0230  EXPT :  -30.0230  ERROR :    0.3100  DIF :    0.0000
DIMETHYL ETHER
CALC :   -47.0051  EXPT :  -43.9720  ERROR :    0.1190  DIF :    2.9141
DIMETHYL OXYMETHANE
CALC :   -81.6772  EXPT :  -83.2140  ERROR :    0.1910  DIF :    1.3458
1,4 DIOXANE
CALC :   -76.4165  EXPT :  -75.4280  ERROR :    0.1910  DIF :    0.7975
FORMALDEHYDE
CALC :   -25.9390  EXPT :  -25.9390  ERROR :    0.1190  DIF :    0.0000
OXALIC ACID
CALC : -172.8530  EXPT : -172.8530  ERROR :    1.1700  DIF :    0.0000
BENZENE
CALC :    19.7530  EXPT :   19.7530  ERROR :    0.1670  DIF :    0.0000
BIPHENYL
CALC :    43.3270  EXPT :   43.3270  ERROR :    0.4780  DIF :    0.0000
MONOBROMO BENZENE
CALC :    25.1740  EXPT :   25.1740  ERROR :    0.9790  DIF :    0.0000
MONOFLUORO BENZENE
CALC :   -27.7060  EXPT :  -27.7060  ERROR :    0.3340  DIF :    0.0000
M DINITRO BENZENE
CALC :    13.3280  EXPT :   13.3280  ERROR :    0.4300  DIF :    0.0000
NAPHTHALENE
CALC :    35.8990  EXPT :   35.8990  ERROR :    0.3580  DIF :    0.0000
ACETONITRILE
CALC :    13.5065  EXPT :   20.9000  ERROR :    0.3000  DIF :    7.0935
PROPIONITRILE
CALC :    11.7070  EXPT :    3.8000  ERROR :    1.0000  DIF :    6.9070
```

[Then following is the Energy Differences data. The same proviso about the number of significant figures as given in the above section on heat of formation data also applies here]

```
 BETWEEN GAUCHE-BUTANE
   AND   ANTI-BUTANE

CALC      0.036 EXPT      0.824 ERROR      0.143 DIF      0.645
```

```
 BETWEEN CIS BUT 2 ENE
   AND   TRANS BUT 2 ENE

CALC      0.138 EXPT      1.130 ERROR      0.380 DIF      0.612

 BETWEEN BOAT-CHAIR BICYCLO[3.3.1] NONANE
   AND   TWIN-CHAIR BICYCLO[3.3.1] NONANE

CALC      3.402 EXPT      3.050 ERROR      0.150 DIF      0.202


 BETWEEN GAUCHE-1,2 DIFLUORO ETHANE
   AND   ANTI 1,2-DIFLUORO ETHANE

CALC      1.182 EXPT      0.770 ERROR      0.400 DIF      0.012

 BETWEEN GAUCHE - 1,1,2,2 TETRA FLUORO ETHANE
   AND   ANTI - 1,1,2,2 TETRA FLUORO ETHANE

CALC      1.182 EXPT      1.170 ERROR      0.200 DIF      0.000

 BETWEEN GAUCHE 1 BROMO PROPANE
   AND   ANTI 1 BROMO PROPANE

CALC      0.591 EXPT      0.110 ERROR      0.030 DIF      0.451

 BETWEEN GAUCHE 1 CHLORO 1 BROMO ETHANE
   AND   ANTI 1 CHLORO 1 -BROMO ETHANE

CALC      1.512 EXPT      1.400 ERROR      0.200 DIF      0.000

 BETWEEN SYN TRANS-2-BUTENOYL CHLORIDE
   AND   ANTI TRANS-2-BUTENOYL CHLORIDE

CALC     -0.012 EXPT      0.100 ERROR      0.200 DIF      0.000

 BETWEEN ANTI-SYN FUMARYL CHLORIDE
   AND   ANTI-ANTI FUMARYL CHLORIDE

CALC     -0.330 EXPT      0.600 ERROR      0.200 DIF      0.730

 BETWEEN SYN-SYN FUMARYL CHLORIDE
   AND   ANTI-ANTI FUMARYL CHLORIDE

CALC     -0.330 EXPT      0.700 ERROR      0.200 DIF      0.830

 BETWEEN CHLORO ACETYL CHLORIDE (GAUCHE)
   AND   CHLORO ACETYL CHLORIDE (ANTI)

CALC     -0.299 EXPT      1.300 ERROR      0.400 DIF      1.199

 BETWEEN GAUCHE 2-BROMO-3-CHLORO-PROPENE
   AND   ANTI 2-BROMO-3-CHLORO PROPENE

CALC      1.444 EXPT      3.000 ERROR      1.000 DIF      0.556
```

```
 BETWEEN GAUCHE 1,2-DIBROMO ETHANE
   AND   ANTI 1,2-DIBROMO ETHANE

CALC     2.048 EXPT     2.200 ERROR     0.140 DIF     0.012

 BETWEEN GAUCHE 1,2-DICHLORO ETHANE
   AND   ANTI 1,2-DICHLORO ETHANE

CALC     1.202 EXPT     1.050 ERROR     0.100 DIF     0.052

 BETWEEN GAUCHE 2,3-DICHLORO-1-PROPENE
   AND   ANTI 2,3-DICHLORO-1-PROPENE

CALC     0.850 EXPT     0.700 ERROR     0.300 DIF     0.000

 BETWEEN GAUCHE 2,3-DICHLORO-1,3-BUTADIENE
   AND   ANTI 2,3-DICHLORO-1,3-BUTADIENE

CALC     2.325 EXPT     2.200 ERROR     1.100 DIF     0.000

 BETWEEN GAUCHE 1,1,2,2-TETRAFLUORO ETHANE
   AND   ANTI 1,1,2,2-TETRAFLUORO ETHANE

CALC     1.214 EXPT     1.170 ERROR     0.100 DIF     0.000
```

As stated above the short duration of the optimisation compared to that for the final optimisation in Chapter 7 means that there is little point in investigating the results on an individual structure basis. In fact due to the size of structures in the database combined with the use of the transputer array for other means during the day and the smaller number of transputers in the network (16 as opposed to 24) it was only possible to do a total of 36 force field loops during the optimisation of the force field for this structure set. This compares with a total of 100 iterations that were done during the final optimisation in the investigation detailed in Chapter 7. Given this restriction it is still possible to examine the general values and to compare these with the same values from the final optimisation of the chapter 7 study. The first comparison is that between the average difference between calculated value and the central experimental value compared to the average of the given errors.

Table 8.2 Comparison of Average Differences Against Errors

Investigation	Chapter 7		This Study	
Interaction Type	Value	Error	Value	Error
Distance	0.0153	0.0224	0.0581	0.0380
Valence Angle	1.6819	1.1635	2.4087	2.0963
Torsion Angle	7.1215	3.7882	5.3090	5.5671
Conformer	0.4446	0.3084	0.5292	0.3084
Heat of Formation	2.519	0.6726	1.2911	0.2499

The units are Ångstrom's, degrees, degrees, Kcal Mol^{-1} and Kcal Mol^{-1} respectively.

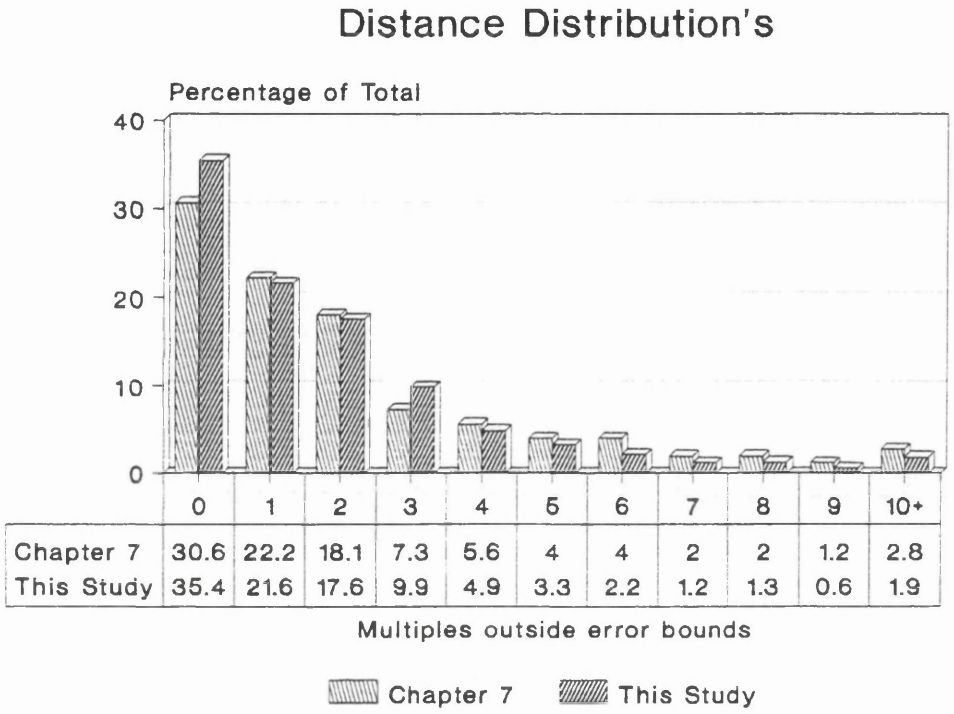
Now if the extension of the force field to a significantly larger number of atom types had resulted in a large increase in the values compared to the errors then it would be a sign that reduction techniques used in determining the parameters could not be used in more general cases.

Proportionally the bond lengths are the worst when compared to the results from chapter 7. It has to be remembered that for this structure set a large number of close non-bonded distances have also been included and that all distances to hydrogen's were included. These points will have two main effects. Firstly it will increase the average error as proportionally more lengths will involve hydrogen's with their associated high errors, especially as normal bonded distances to hydrogen's were less common in electron diffraction studies. Secondly the non-bonded distances are more likely to be further from their experimental values as their relative positions are affected by the deviation of all the atoms between them.

Considering these points the values for the distances could be considered as reasonable and not to indicate any large errors in the force field. This is to be expected as the bond lengths are amongst the first to settle down to near their final values and the parameters used to calculate the bond stretch force constants may optimise considerably slower but have a far smaller effect on the calculated bond lengths. This assumption is confirmed by the plot in figure 8.7. This compares, in percentage terms, the distribution of the distances for the final optimisations in both this study and that done for the study in Chapter 7, where the final optimisation went through 100 force field loops. This graph shows that the average difference between the percentages in all columns is only 1.33% with this study having 4.8% more of the distances in the 'within error' band and a lower percentage in all those error bands with a 'times error' rating of greater than 3.

For the valence angles the ratio's of average error to the difference between experimental and calculated for both cases are similarly close. Figure 8.9 shows, again in percentage terms, the distribution of the valence angle residual factors for this study and for the final optimisation in Chapter 7. This shows that these distributions are even closer than those for the distances, with an average difference between the two cases of only 1.13%. For this measure the Chapter 7 study has a slightly better distribution, especially in the '10+' category, which is likely to be responsible for producing a slightly worse standard deviation for this study.

Figures 8.7 and 8.8 Correlation Distribution Comparisons Between Studies for Distances and Torsion angles



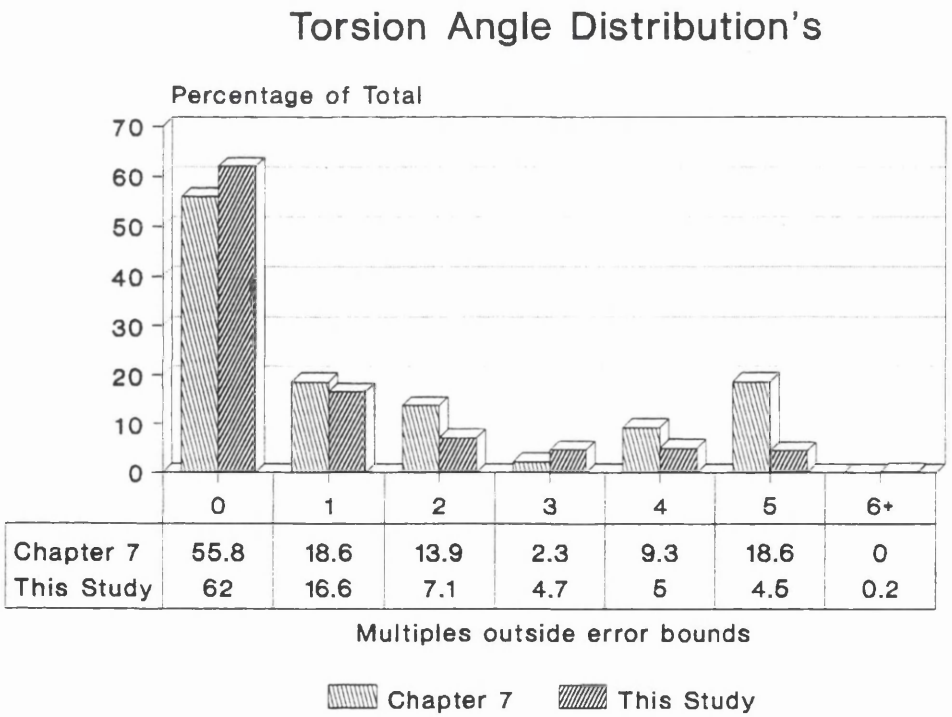
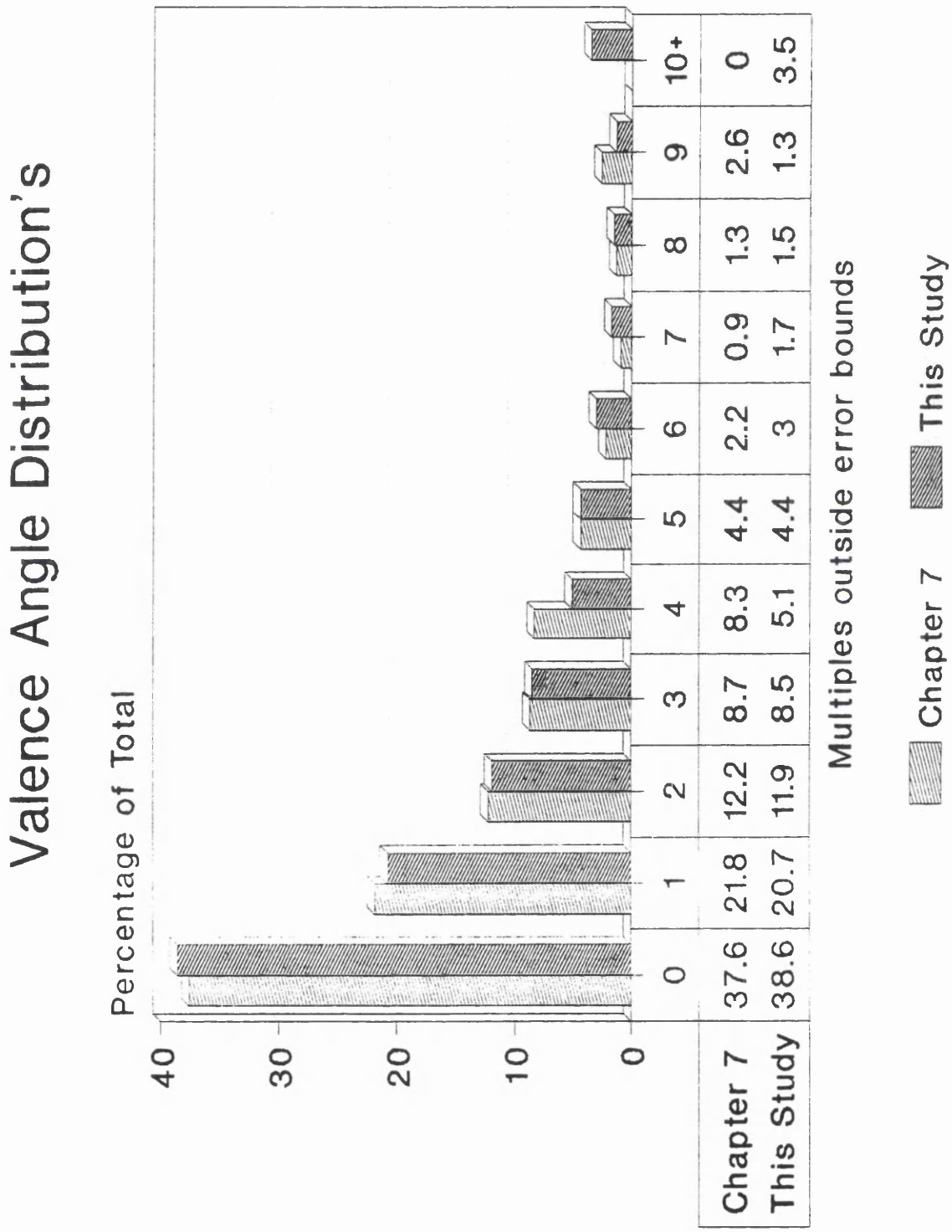
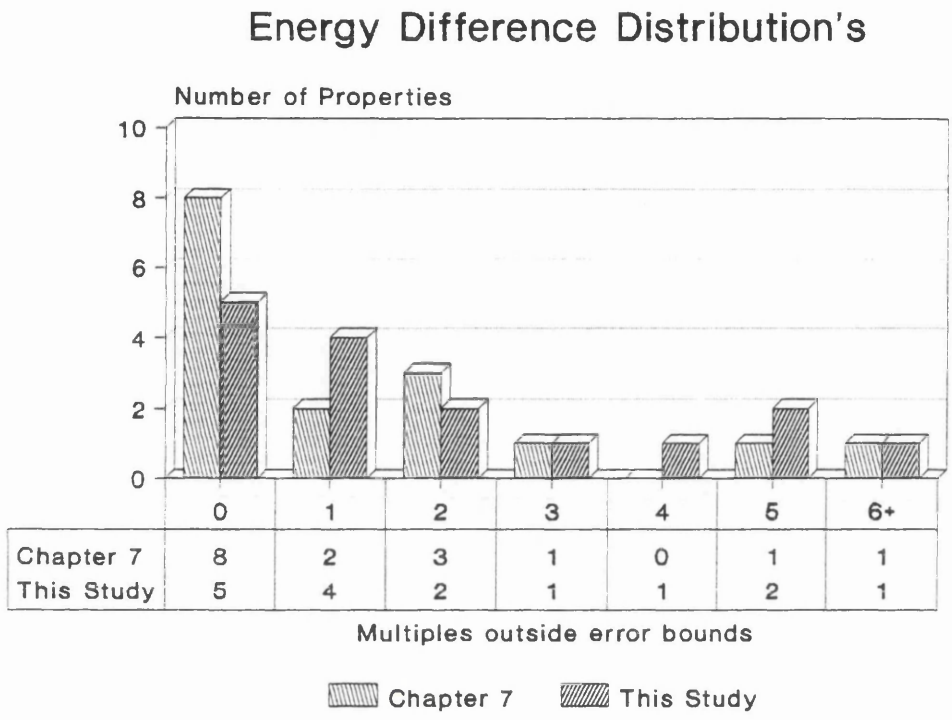


Figure 8.9 Distribution Comparison for Valence Angle Residual Factors



Figures 8.10 and 8.11 Residual Factor Distribution Comparisons for Energy Measures



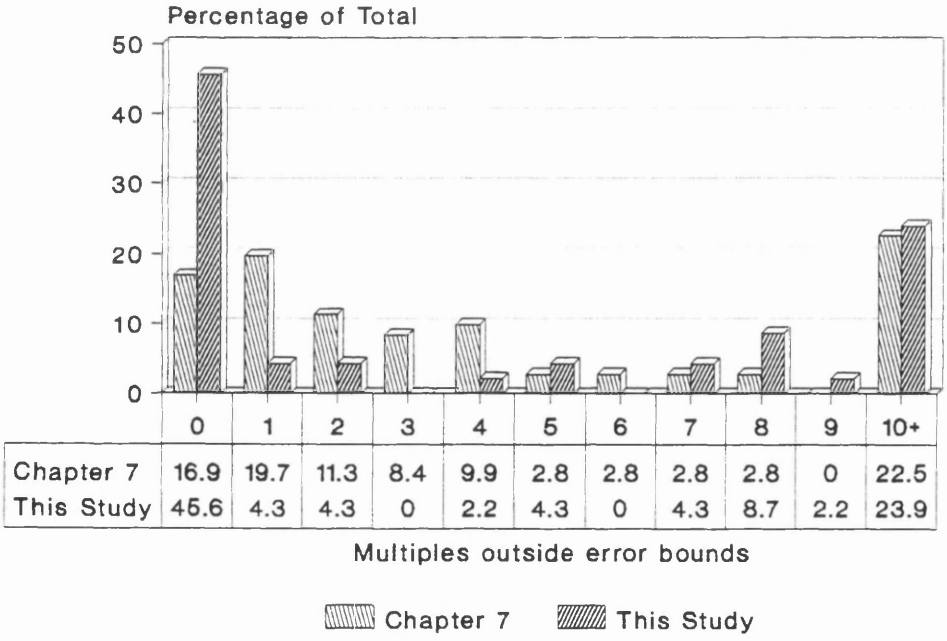


Figure 8.8 shows a similar plot for the torsion angles. Despite a large increase in the number of torsion angles from 51 to 3330 the distributions are again close. This is slightly misleading as the torsion angles in the Chapter 7 study were those that were considered worth mentioning by the original authors. This was usually when the final value of the torsion angle arose from a balance of several forces. Whereas for this study all torsion angles were used, including all those where the effect of all other forces outside of the torsional barrier are minimal. For example the torsion angle in ethane would not have been used in the Chapter 7 study, whereas in this study it would be included.

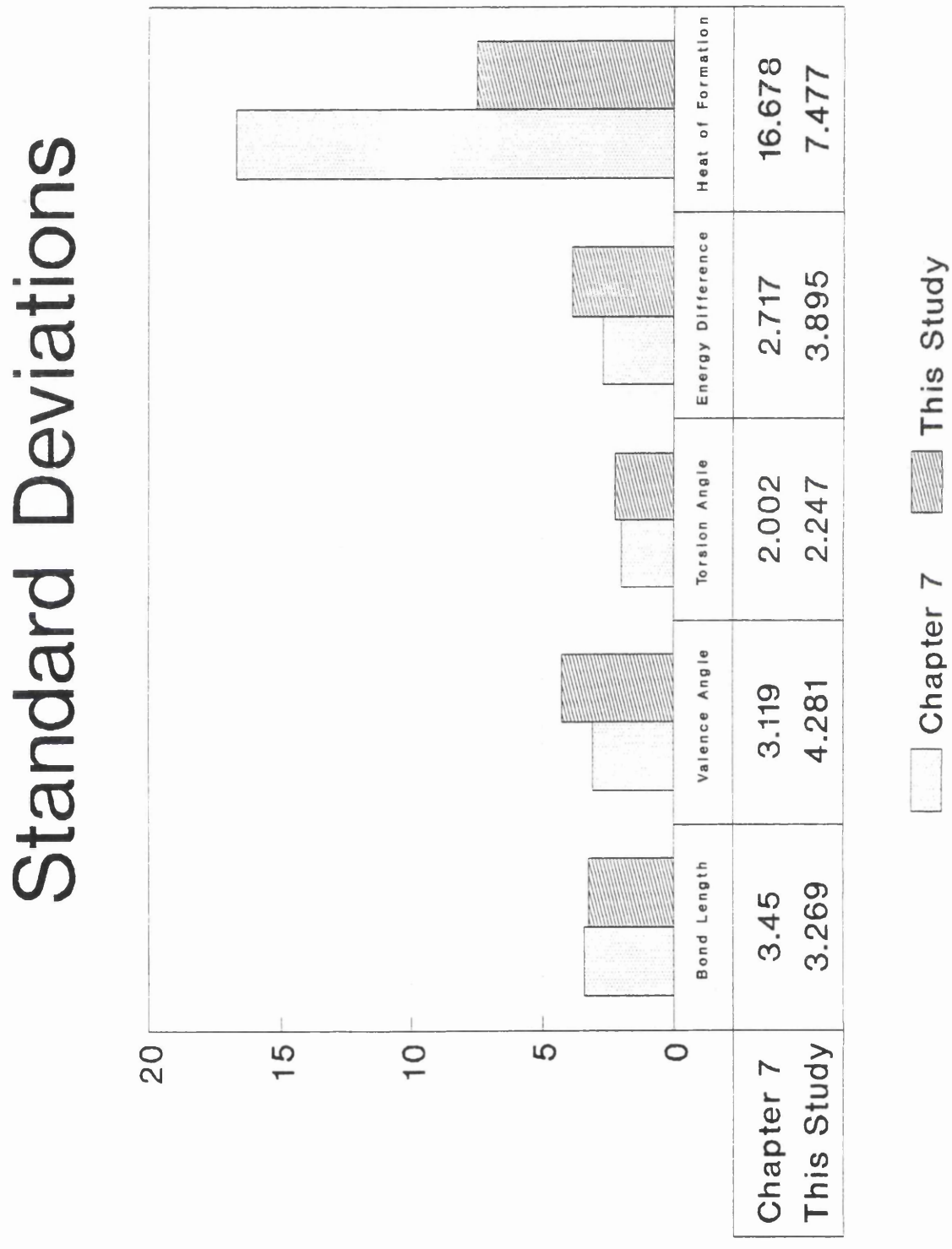
Figure 8.10 shows the comparison for the energy difference factors. This is the only case where a direct comparison is possible and it can be seen that these measures have suffered at the expense of the structural properties. Due to the limited number of properties there is little that can be deduced from this graph.

Figure 8.11 shows the distributions for the heat of formation residual factors. The distribution of the residual factors for this study has been commented on in the above section, the point to notice from this graph is the significantly greater number of structures that are in the 'within error' band. Why this should happen is not obvious, especially as the overall residual factor for the heat of formation residual factors got worse over the course of the optimisation. It is most likely that it is related to the large increase in the number of geometric enthalpy increments that are required for the larger number of atom types.

Figure 8.12 shows the comparison of the standard deviations for these interaction types for the two studies, note that the value used for distance for this study is for the bond distances rather than the non-bonded interactions as the Chapter 7 study has a minimal number of non-bonded distances.

The distance residual factor is slightly lower in this study, almost certainly because the spread of electronegativities is not a pronounced as the Chapter 7 study. This will lead to less deviation for the average bond, thus the lower residual factors. For the valence angle residual factor the Chapter 7 Study comes off significantly better, due almost entirely to the very few cases with exceptionally high residual factors in this study. For the Torsion Angles the numbers reflect the discussion above on figure 8.8. Similarly the increase in the standard deviation for the Energy Difference residual factors is as would be expected from the distributions. The largest change is in the Heat of Formation residual factors, what this shows is that not only are the low residual factors better but that the few cases with very large residual factors are also significantly better in this study.

Figure 8.12 Standard Deviations Comparisons Between Investigations



8.5.3 Final Parameter Values

The following tables contain the parameters at the end of the final optimisation. These values will be discussed in the section that follows on from these tables. When examining the parameters the following notes should be taken into account.

Bond Lengths

All bond lengths that are given to three decimal places are those lengths that have been taken from CHEMMIN[71] and which have not been optimised in this study. The list associated numbers in brackets is the bond stretch force constants calculated using the final parameter values for the bond stretch calculation. The three cases where the bond strength is given as '*****' are those cases where the negative factor in the calculation equation becomes large enough to make the total negative. Luckily none of these bonds were involved in this study and it would be necessary to put some lower limit on the calculated bond strength before this force field could be used when these bond types were present. As stated in Chapter 7 the number of significant figures given for those parameters that have been optimised is the same as that with which these parameters had been saved. These are given to this degree of accuracy to maximise the correlation between the parameters and the optimised results.

Strain free Angle

In this section there are two additional parameters. The 'Lone Pair Distance' is the equivalent to a bond length and is used when calculating the angles for those atom types where a lone pair in involved, e.g. N_{sp3}. The second term is a 'H Factor' which is the angle measure which was subtracted from the total angle for each hydrogen attached to the central atom. This was initially introduced when a study of the strain free angles in force fields such as the WBFF[72] showed that the total of the various strain free angles varied with the number of hydrogen's. However, the optimised values tended to become far greater than the values deduced from the other force fields and so this feature was removed before the final study in Chapter 7. Note also that the N_{sp} has not been optimised. This is expected as this is a linear configuration so the default strain free angle will always be 180° and this anticipated within OPTPARAM.

Bond Lengths

All lengths in Ångstroms. Number in brackets is calculated bond stretch force constant in Kcal Mol^{-1} $Å^{-2}$.

H	- H	0.746	(697.4)
H	- C_{sp}	1.075651	(335.5)
H	- C_{sp2}	1.068077	(340.2)
H	- C_{arom}	1.083633	(330.5)
H	- C_{sp3}	1.093432	(324.6)
H	- N_{sp2}	0.990	(396.0)
H	- N_{amide}	1.025194	(369.3)
H	- N_{sp3}	1.021866	(371.7)
H	- N_{cation}	0.878563	(502.8)
H	- O_{sp3}	0.971279	(411.4)
H	- O_{anion}	0.970	(412.5)
H	- F	0.920	(458.6)
H	- P_{iii}	1.430	(189.8)
H	- P_{v}	1.348711	(213.4)
H	- S_{-ii-}	1.440586	(187.0)
H	- S_{iv}	1.320	(222.8)
H	- S_{vi}	1.320	(222.8)
H	- Cl	1.270	(240.6)
H	- Br	1.410	(195.2)
H	- I	1.610	(149.7)
C_{sp}	- C_{sp}	1.205408	(869.7)
C_{sp}	- C_{sp2}	1.466702	(410.4)
C_{sp}	- C_{arom}	1.457861	(419.3)
C_{sp}	- C_{sp3}	1.461030	(416.0)
C_{sp}	- N_{sp}	1.158674	(1079.1)
C_{sp}	- N_{sp2}	1.309577	(616.5)
C_{sp}	- N_{sp3}	1.270	(693.9)
C_{sp}	- N_{cation}	1.200	(888.9)
C_{sp}	- O_{sp3}	1.230	(792.7)
C_{sp}	- O_{anion}	1.210	(854.1)
C_{sp}	- F	1.260	(716.2)
C_{sp}	- P_{iii}	1.760	(207.0)
C_{sp}	- P_{v}	1.780	(197.6)
C_{sp}	- S_{-ii-}	1.678876	(249.8)
C_{sp}	- Cl	1.630	(279.6)
C_{sp}	- Br	1.797391	(189.7)
C_{sp}	- I	1.999906	(115.7)
C_{sp2}	- C_{sp2}	1.344154	(560.0)
C_{sp2}	- C_{arom}	1.491513	(386.6)
C_{sp2}	- C_{sp3}	1.511777	(368.4)
C_{sp2}	- N_{sp2}	1.342123	(563.1)
C_{sp2}	- N_{arom}	1.480	(397.4)
C_{sp2}	- N_{amide}	1.351445	(549.2)
C_{sp2}	- N_{nitro}	1.395493	(489.5)
C_{sp2}	- N_{sp3}	1.376237	(514.4)
C_{sp2}	- N_{cation}	1.480	(397.4)
C_{sp2}	- O_{sp2}	1.209689	(855.1)
C_{sp2}	- O_{sp3}	1.338525	(568.7)
C_{sp2}	- O_{anion}	1.261639	(712.5)
C_{sp2}	- F	1.320135	(598.3)
C_{sp2}	- P_{iii}	1.770	(202.3)
C_{sp2}	- P_{v}	1.730	(222.0)
C_{sp2}	- S_{-ii-}	1.763024	(205.6)
C_{sp2}	- $S_{ii=}$	1.653855	(264.6)
C_{sp2}	- S_{iv}	1.710	(232.5)
C_{sp2}	- S_{vi}	1.612140	(291.4)
C_{sp2}	- Cl	1.700379	(237.7)
C_{sp2}	- Br	1.804111	(186.7)
C_{sp2}	- I	2.035052	(105.5)
C_{arom}	- C_{arom}	1.372534	(519.4)
C_{arom}	- C_{sp3}	1.520000	(361.3)
C_{arom}	- N_{sp2}	1.370291	(522.4)
C_{arom}	- N_{arom}	1.333229	(577.0)
C_{arom}	- N_{amide}	1.365544	(529.0)
C_{arom}	- N_{nitro}	1.452164	(425.1)
C_{arom}	- N_{sp3}	1.370360	(522.4)
C_{arom}	- N_{cation}	1.447769	(429.7)
C_{arom}	- O_{sp2}	1.230	(792.7)
C_{arom}	- O_{sp3}	1.370532	(522.1)
C_{arom}	- O_{anion}	1.290	(652.8)
C_{arom}	- F	1.336906	(571.2)
C_{arom}	- P_{iii}	1.760	(207.0)
C_{arom}	- P_{v}	1.780	(197.6)
C_{arom}	- S_{-ii-}	1.761476	(206.3)
C_{arom}	- S_{iv}	1.760	(207.0)
C_{arom}	- S_{vi}	1.740871	(216.4)
C_{arom}	- Cl	1.717105	(228.7)
C_{arom}	- Br	1.863981	(162.0)
C_{arom}	- I	2.093616	(90.0)
C_{sp3}	- C_{sp3}	1.537741	(346.5)

C_{sp3}	- N_{sp2}	1.471902	(405.2)
C_{sp3}	- N_{arom}	1.480	(397.4)
C_{sp3}	- N_{amide}	1.455550	(421.6)
C_{sp3}	- N_{nitro}	1.479685	(397.7)
C_{sp3}	- N_{sp3}	1.472972	(404.2)
C_{sp3}	- N_{cation}	1.486775	(391.0)
C_{sp3}	- O_{sp3}	1.417725	(462.8)
C_{sp3}	- O_{anion}	1.380	(509.4)
C_{sp3}	- F	1.365689	(528.8)
C_{sp3}	- P_{iii}	1.828637	(176.2)
C_{sp3}	- P_{v}	1.848985	(167.9)
C_{sp3}	- S_{-ii-}	1.822876	(178.7)
C_{sp3}	- S_{iv}	1.820057	(179.9)
C_{sp3}	- S_{vi}	1.773064	(200.8)
C_{sp3}	- Cl	1.776731	(199.1)
C_{sp3}	- Br	1.964303	(126.6)
C_{sp3}	- I	2.144590	(77.7)
N_{sp}	- N_{sp}	1.120	(1393.8)
N_{sp2}	- N_{sp2}	1.312285	(611.8)
N_{sp2}	- N_{amide}	1.352729	(547.3)
N_{sp2}	- N_{nitro}	1.374286	(517.0)
N_{sp2}	- N_{sp3}	1.409557	(472.4)
N_{sp2}	- O_{sp2}	1.235769	(776.8)
N_{sp2}	- O_{sp3}	1.435270	(443.1)
N_{sp2}	- O_{anion}	1.250	(740.0)
N_{sp2}	- F	1.360	(536.8)
N_{sp2}	- P_{iii}	1.620	(286.2)
N_{sp2}	- P_{v}	1.583820	(311.2)
N_{sp2}	- S_{-ii-}	1.694543	(240.9)
N_{sp2}	- $S_{ii=}$	1.530	(352.9)
N_{sp2}	- S_{iv}	1.530	(352.9)
N_{sp2}	- S_{vi}	1.530	(352.9)
N_{sp2}	- Cl	1.791175	(192.5)
N_{sp2}	- Br	2.139866	(78.8)
N_{arom}	- N_{arom}	1.330	(582.1)
N_{amide}	- N_{amide}	1.384300	(503.8)
N_{amide}	- N_{nitro}	1.394282	(491.1)
N_{amide}	- N_{sp3}	1.420	(460.2)
N_{amide}	- P_{v}	1.700	(237.9)
N_{amide}	- S_{vi}	1.621670	(285.1)
N_{amide}	- B_{r}	1.820	(179.9)
N_{nitro}	- N_{sp3}	1.340	(566.4)
N_{nitro}	- O_{sp2}	1.226340	(803.2)
N_{nitro}	- O_{sp3}	1.452883	(424.4)
N_{nitro}	- O_{anion}	1.220	(822.2)
N_{sp3}	- N_{sp3}	1.398108	(486.3)
N_{sp3}	- N_{cation}	1.407788	(474.5)
N_{sp3}	- O_{sp3}	1.451419	(425.9)
N_{sp3}	- O_{anion}	1.270	(693.9)
N_{sp3}	- F	1.361398	(534.8)
N_{sp3}	- P_{iii}	1.670	(255.0)
N_{sp3}	- P_{v}	1.680	(249.2)
N_{sp3}	- S_{-ii-}	1.725810	(224.1)
N_{sp3}	- S_{iv}	1.620	(286.2)
N_{sp3}	- S_{vi}	1.623337	(284.0)
N_{sp3}	- Cl	1.750	(211.9)
N_{sp3}	- Br	2.140	(78.8)
N_{sp3}	- I	2.300	(46.5)
N_{cation}	- N_{cation}	1.270	(693.9)
N_{cation}	- O_{sp3}	1.200	(888.9)
N_{cation}	- O_{anion}	1.400	(484.0)
N_{cation}	- F	1.360	(536.8)
N_{cation}	- P_{iii}	1.670	(255.0)
N_{cation}	- S_{-ii-}	1.620	(286.2)
N_{cation}	- S_{iv}	1.620	(286.2)
N_{cation}	- S_{vi}	1.767466	(203.5)
N_{cation}	- Cl	1.750	(211.9)
N_{cation}	- Br	2.140	(78.8)
N_{cation}	- I	2.220	(61.5)
O_{sp2}	- O_{sp2}	1.210	(854.1)
O_{sp2}	- P_{v}	1.440815	(437.1)
O_{sp2}	- $S_{ii=}$	1.410	(471.9)
O_{sp2}	- S_{iv}	1.487318	(390.5)
O_{sp2}	- S_{vi}	1.437982	(440.2)
O_{sp2}	- Cl	1.470	(407.1)
O_{sp2}	- Br	1.610	(292.9)
O_{sp2}	- I	1.800	(188.6)
O_{sp3}	- O_{sp3}	1.486875	(390.9)
O_{sp3}	- O_{anion}	1.260	(716.2)
O_{sp3}	- F	1.418	(462.5)
O_{sp3}	- P_{iii}	1.500	(378.9)
O_{sp3}	- P_{v}	1.573464	(318.8)
O_{sp3}	- S_{-ii-}	1.430	(448.9)
O_{sp3}	- S_{iv}	1.620	(286.2)

O_{sp3}	- S_{vi}	1.621466	(285.2)
O_{sp3}	- Cl	1.673659	(252.8)
O_{sp3}	- Br	1.850	(167.5)
O_{sp3}	- I	2.000	(115.6)
O_{anion}	- P_{v}	1.505530	(373.9)
O_{anion}	- S_{-ii-}	1.310	(615.8)
O_{anion}	- S_{iv}	1.510	(370.0)
O_{anion}	- S_{vi}	1.460557	(416.5)
O_{anion}	- Cl	1.560	(328.9)
O_{anion}	- Br	1.630	(279.6)
O_{anion}	- I	1.770	(202.3)
F	- F	1.417	(463.7)
F	- P_{iii}	1.535	(348.8)
F	- P_{v}	1.543911	(341.5)
F	- S_{-ii-}	1.600406	(299.4)
F	- S_{iv}	1.585	(310.3)
F	- S_{vi}	1.582150	(312.4)
F	- Cl	1.630	(279.6)
F	- Br	1.760	(207.0)
F	- I	1.860	(163.6)
P_{iii}	- P_{iii}	2.200	(65.6)
P_{iii}	- P_{v}	2.210	(63.5)
P_{iii}	- S_{-ii-}	1.860	(163.6)
P_{iii}	- S_{iv}	2.000	(115.6)
P_{iii}	- S_{vi}	2.000	(115.6)
P_{iii}	- Cl	2.100	(88.4)
P_{iii}	- Br	2.130	(81.1)
P_{v}	- P_{v}	2.210	(63.5)
P_{v}	- S_{-ii-}	2.140	(78.8)
P_{v}	- $S_{ii=}$	1.881102	(155.5)
P_{v}	- Cl	1.998222	(116.2)
P_{v}	- Br	2.171951	(71.6)
P_{v}	- I	2.660	(*****)
S_{-ii-}	- S_{-ii-}	2.030885	(106.7)
S_{-ii-}	- S_{iv}	2.120	(83.5)
S_{-ii-}	- S_{vi}	2.120	(83.5)
S_{-ii-}	- Cl	2.023858	(108.7)
S_{-ii-}	- Br	2.270	(51.9)
S_{-ii-}	- I	2.630	(1.6)
$S_{ii=}$	- $S_{ii=}$	2.060	(98.7)
S_{iv}	- S_{iv}	2.120	(83.5)
S_{iv}	- Cl	2.330	(41.3)
S_{iv}	- Br	2.450	(23.1)
S_{vi}	- S_{vi}	2.070	(96.0)
S_{vi}	- Cl	2.066082	(97.1)
S_{vi}	- Br	2.270	(51.9)
Cl	- Cl	1.988	(119.3)
Cl	- Br	2.140	(78.8)
Cl	- I	2.320	(43.0)
Br	- Br	2.290	(48.2)
Br	- I	2.670	(*****)
I	- I	2.820	(*****)

Conjugated Bond Lengths

Csp	- Csp	1.364863	(529.9)
Csp	- Csp2	1.420	(460.2)
Csp	- Carom	1.370	(522.8)
Csp2	- Csp2	1.477310	(400.0)
Csp2	- Carom	1.470	(407.1)
Csp2	- Nsp2	1.430	(448.9)
Csp2	- Narom	1.480	(397.4)
Csp2	- Namide	1.415131	(465.9)
Csp2	- Pv	1.790	(193.0)
Csp2	- Siv	1.820	(179.9)
Csp2	- Svi	1.839604	(171.7)
Carom	- Carom	1.500	(378.9)
Carom	- Pv	1.440	(438.0)
Carom	- Nsp2	1.420	(460.2)
Carom	- Narom	1.440	(438.0)
Carom	- Namide	1.440	(438.0)
Carom	- Siv	1.760	(207.0)
Carom	- Svi	1.740	(216.9)
Nsp2	- Nsp2	1.390	(496.5)
Nsp2	- Namide	1.390	(496.5)
Nsp2	- Pv	1.600	(299.7)
Nsp2	- Svi	1.650	(267.0)
Namide	- Namide	1.390	(496.5)
Namide	- Pv	1.700	(237.9)
Namide	- Svi	1.640	(273.2)
Pv	- Pv	2.220	(61.5)
Svi	- Svi	2.190	(67.7)

Bond Stretch Calculation Values

$$½k_r(ij)=\left[\frac{c_1}{l_o(ij)^2}\right]+\left[\frac{c_2}{(l_o(ij)-1)^2}\right]-\left[\frac{c_3}{l_o(ij)}\right]$$

$$½k_r(iH)=\frac{c_4}{l_o(iH)^2}$$

C_1 1828.577759 Kcal Mol^{-1}

C_2 8.068143 Kcal Mol^{-1}

C_3 699.170105 Kcal Mol^{-1} $Å^{-1}$

C_4 388.136108 Kcal Mol^{-1}

Strain Free Angles

A) weights

Unitless

H	108.505547
Lone Pair	101.048508
C_{sp}	136.970215
C_{sp2}	133.034119
C_{arom}	134.074860
C_{sp3}	139.746719
N_{sp}	130.000
N_{sp2}	133.089066
N_{arom}	130.428040
N_{amide}	136.092682
N_{nitro}	125.913620
N_{sp3}	134.171188
N_{cation}	128.175186
O_{sp2}	141.436920
O_{sp3}	132.422562
O_{anion}	133.213165
F	119.015396
P_{iii}	173.947678
P_v	156.132874
S_{-ii-}	150.163986
$S_{ii=}$	172.389877
S_{iv}	148.535507
S_{vi}	145.877853
Cl	137.768890
Br	133.110809
I	151.495163

B) Total Angles

C_{sp2}	393.690521°
C_{arom}	350.105713°
C_{sp3}	661.219910°
N_{sp2}	340.819214°
N_{arom}	321.568085°
N_{amide}	369.393097°
N_{nitro}	362.920898°
N_{sp3}	657.859497°
N_{cation}	699.252747°
O_{sp3}	650.970825°
P_{iii}	626.332764°
P_v	660.255066°
S_{-ii-}	614.219055°
S_{iv}	637.630371°
S_{vi}	879.958923°

C) Other Values

Exponent	0.307813
H Factor	1.215879°
Lone Pair Distance	0.944561Å

Angle Bend Calculation Values

$$½k_\theta(ijk)=\frac{d_1\left\{d_2+d_3\left(|X(i)-X(j)|+|X(j)-X(k)|\right)\right\}}{l_o(ij)l_o(jk)}$$

$$½k_\theta(ijH)=\frac{d_4\times d_1\left\{d_2+d_3\left(|X(i)-X(j)|+|X(j)-X(H)|\right)\right\}}{l_o(ij)l_o(jH)}$$

$$½k_\theta(HjH)=\frac{d_5\times d_1\left\{d_2+d_3\left(|X(H)-X(j)|+|X(j)-X(H)|\right)\right\}}{l_o(Hj)l_o(jH)}$$

d1 0.001578 Kcal Mol^{-1} Deg^{-2} $Å^2$

d2 15.953610

d3 2.277630

d4 0.450614

d5 0.205958

Non-Bonded Parameters

A) VdW Radii

All in Ångstroms

H	1.187504
C_{sp}	1.656939
C_{sp2}	1.016964
C_{arom}	1.957636
C_{sp3}	1.704498
N_{sp}	2.287631
N_{sp2}	1.173209
N_{arom}	1.754929
N_{amide}	1.511304
N_{nitro}	1.560281
N_{sp3}	1.183461
N_{cation}	2.180137
O_{sp2}	1.344306
O_{sp3}	1.440926
O_{anion}	1.878716
F	1.419256
P_{iii}	1.443697
P_{v}	1.967908
S_{-ii-}	0.792987
$S_{ii=}$	2.004262
S_{iv}	1.229532
S_{vi}	2.512208
Cl	1.892437
Br	2.152779
I	2.211423

B) *f* Factor

All Kcal $Mole^{-1}$

H	0.041173
C_{sp}	0.124287
C_{sp2}	0.133295
C_{arom}	0.108029
C_{sp3}	0.103464
N_{sp}	0.118060
N_{sp2}	0.091026
N_{arom}	0.107093
N_{amide}	0.096567
N_{nitro}	0.097175
N_{sp3}	0.098281
N_{cation}	0.105499
O_{sp2}	0.122681
O_{sp3}	0.113077
O_{anion}	0.123366
F	0.105131
P_{iii}	0.396558
P_{v}	0.420421
S_{-ii-}	0.265928
$S_{ii=}$	0.183546
S_{iv}	0.278885
S_{vi}	0.317016
Cl	0.292708
Br	0.313988
I	0.604453

Torsional Barriers

All Kcal Mol^{-1}

C_{sp}	-	C_{sp2}	17.715
C_{sp}	-	C_{arom}	0.275
C_{sp}	-	C_{sp3}	0.011
C_{sp}	-	N_{sp2}	23.944
C_{sp}	-	N_{sp3}	0.011
C_{sp}	-	N_{cation}	0.011
C_{sp}	-	O_{sp3}	0.011
C_{sp}	-	P_{iii}	0.011
C_{sp}	-	P_{v}	17.715
C_{sp}	-	S_{-ii-}	0.008
C_{sp2}	-	C_{sp2}	8.125
C_{sp2}	-	C_{arom}	0.708
C_{sp2}	-	C_{sp3}	0.182
C_{sp2}	-	N_{sp2}	2.250
C_{sp2}	-	N_{arom}	0.708
C_{sp2}	-	N_{amide}	3.417
C_{sp2}	-	N_{nitro}	0.463
C_{sp2}	-	N_{sp3}	4.889
C_{sp2}	-	N_{cation}	4.889
C_{sp2}	-	O_{sp3}	2.725
C_{sp2}	-	P_{iii}	1.100
C_{sp2}	-	P_{v}	8.125
C_{sp2}	-	S_{-ii-}	0.475

C_{sp2}	S_{iv}	6.000	N_{arom}	P_{iii}	0.008
C_{sp2}	S_{vi}	0.097	N_{amide}	N_{amide}	1.205
C_{arom}	C_{arom}	0.708	N_{amide}	N_{nitro}	1.000
C_{arom}	C_{sp3}	0.182	N_{amide}	P_{v}	0.889
C_{arom}	N_{sp2}	2.250	N_{amide}	S_{vi}	0.233
C_{arom}	N_{arom}	0.708	N_{nitro}	N_{sp3}	0.006
C_{arom}	N_{amide}	0.708	N_{nitro}	O_{sp3}	5.000
C_{arom}	N_{nitro}	0.588	N_{sp3}	N_{sp3}	1.205
C_{arom}	N_{sp3}	0.825	N_{sp3}	N_{cation}	1.205
C_{arom}	N_{cation}	0.825	N_{sp3}	O_{sp3}	2.475
C_{arom}	O_{sp3}	0.821	N_{sp3}	P_{iii}	1.000
C_{arom}	P_{iii}	0.008	N_{sp3}	P_{v}	4.889
C_{arom}	P_{v}	0.708	N_{sp3}	S_{-ii-}	1.667
C_{arom}	S_{-ii-}	0.190	N_{sp3}	S_{iv}	1.667
C_{arom}	S_{iv}	0.500	N_{sp3}	S_{vi}	0.233
C_{arom}	S_{vi}	0.100	N_{sp3}	S_{cation}	0.620
C_{sp3}	C_{sp3}	0.133	N_{cation}	N_{cation}	1.205
C_{sp3}	N_{sp2}	0.083	N_{cation}	O_{sp3}	0.215
C_{sp3}	N_{arom}	0.008	N_{cation}	P_{iii}	1.000
C_{sp3}	N_{amide}	0.083	N_{cation}	S_{-ii-}	1.667
C_{sp3}	N_{nitro}	0.006	N_{cation}	S_{iv}	1.667
C_{sp3}	N_{sp3}	0.114	N_{cation}	S_{vi}	0.233
C_{sp3}	N_{cation}	0.114	O_{sp3}	O_{sp3}	3.500
C_{sp3}	O_{sp3}	0.100	O_{sp3}	P_{iii}	0.163
C_{sp3}	P_{iii}	0.163	O_{sp3}	P_{v}	2.725
C_{sp3}	P_{v}	0.389	O_{sp3}	S_{-ii-}	0.400
C_{sp3}	S_{-ii-}	0.195	O_{sp3}	S_{iv}	0.400
C_{sp3}	S_{iv}	0.245	O_{sp3}	S_{vi}	0.310
C_{sp3}	S_{vi}	0.245	P_{iii}	P_{iii}	0.513
N_{sp}	N_{sp2}	1.680	P_{iii}	P_{v}	0.071
N_{sp2}	N_{sp2}	40.000	P_{iii}	S_{-ii-}	0.100
N_{sp2}	N_{amide}	0.783	P_{iii}	S_{iv}	0.100
N_{sp2}	N_{nitro}	0.625	P_{iii}	S_{vi}	0.100
N_{sp2}	N_{sp3}	0.783	P_{v}	P_{v}	4.065
N_{sp2}	O_{sp3}	5.000	P_{v}	S_{-ii-}	5.000
N_{sp2}	P_{iii}	0.083	S_{-ii-}	S_{-ii-}	4.070
N_{sp2}	P_{v}	45.000	S_{-ii-}	S_{iv}	4.070
N_{sp2}	S_{-ii-}	4.300	S_{-ii-}	S_{vi}	4.070
N_{sp2}	S_{iv}	60.000	S_{iv}	S_{iv}	10.000
N_{sp2}	S_{vi}	0.467	S_{vi}	S_{vi}	10.000
N_{arom}	N_{arom}	5.000			
N_{arom}	O_{sp3}	0.008			

Conjugated Torsion Barriers

C_{sp}	- C_{sp}	1.000
C_{sp}	- C_{sp2}	0.750
C_{sp}	- C_{arom}	0.750
C_{sp}	- P_{v}	0.400
C_{sp2}	- C_{sp2}	0.538
C_{sp2}	- C_{arom}	0.568
C_{sp2}	- N_{sp2}	0.463
C_{sp2}	- N_{arom}	0.588
C_{sp2}	- S_{iv}	0.425
C_{sp2}	- S_{vi}	0.425
C_{arom}	- C_{arom}	0.250
C_{arom}	- N_{sp}	0.275
C_{arom}	- N_{sp2}	0.588
C_{arom}	- N_{arom}	0.250
C_{arom}	- N_{amide}	0.250
C_{arom}	- P_{v}	0.300
C_{arom}	- S_{iv}	0.425
C_{arom}	- S_{vi}	0.425
N_{sp2}	- N_{sp2}	0.625
N_{sp2}	- P_{v}	0.230
P_{v}	- P_{v}	0.270
S_{vi}	- S_{vi}	0.350

Out-of-Plane Bending Force Constants

All Kcal Mole^{-1} Deg^{-2}

C_{sp2}	0.001208
C_{arom}	0.001160
N_{amide}	0.000189
N_{nitro}	0.000213

8.5.3.1 Bond Lengths

The following table 8.3 contains a selection of bond lengths for this study, that in chapter 7, Cosmic and CHEMMIN.

Table 8.3 Comparison of Bond Length Parameters

Bond	This Study	Chapter 7	Cosmic	CHEMMIN
C_{sp2}-H	1.0681	1.0860	1.089	1.070
C_{sp3}-H	1.0934	1.0950	1.100	1.091
C_{sp2}=C_{sp2}	1.3442	1.3250	1.335	1.335
C_{sp2}-C_{sp2}*	1.4773	1.4453	------	1.480
C_{sp2}-C_{sp3}	1.5118	1.4960	1.501	1.520
C_{sp2}=O_{sp2}	1.2097	1.1960	1.220	1.210
C_{sp2}-F	1.3201	1.3270	1.330	1.330
C_{sp2}-Cl	1.7004	1.7425	1.750	1.720
C_{sp2}-Br	1.8041	1.9114	------	1.890
C_{sp3}-C_{sp3}	1.5377	1.4872	1.520	1.541
C_{sp3}-F	1.3657	1.3407	1.360	1.381
C_{sp3}-Cl	1.7767	1.7670	1.767	1.769
C_{sp3}-Br	1.9643	1.9233	1.938	1.937

* Conjugated bond.

There are no values for this study which show any major variation from those taken from the two force fields. The parameter showing the greatest difference is that for the C_{sp2}-Br bond which is significantly shorter than the CHEMMIN value. This is almost certainly due to the under representation of bromine in the structure set such that the selection of such distances is not representative. Otherwise all other bonds are within 0.02Å of one or both of the other force field parameters.

8.5.3.2 Bond Stretch Force Constant Calculation Values

These values are much closer to those from CHEMMIN from which the initial values were originally taken and consequently the calculated values are considerably closer to those found in other force field. As an example of this table 8.4 is a copy of the table of bond strengths taken from Chapter 7 but with the bond stretch values from this study added. All bond stretch force constants given have the units of Kcal Mol^{-1} $Å^{-2}$.

Table 8.4 Comparison of Bond Stretch Force Constants.

Bond	This Study	Chapter 7	Cosmic	CHEMMIN
C_{sp2}-H	340.2	244.2	346.0	334.9
C_{sp3}-H	324.6	240.2	331.2	329.4
C_{sp2}=C_{sp2}	560.0	312.9	670.0	594.4
C_{sp2}-C_{sp2}*	400.0	204.4	------	438.0
C_{sp2}-C_{sp3}	368.4	182.9	319.5	407.1
C_{sp2}=O_{sp2}	855.1	519.9	777.6	903.6
C_{sp2}-F	598.3	310.6	600.0	591.1
C_{sp2}-Cl	237.7	55.7	260.0	222.6
C_{sp2}-Br	186.7	8.3	------	151.7
C_{sp3}-C_{sp3}	346.5	175.5	316.8	396.6
C_{sp3}-F	528.8	295.9	300.0	569.7
C_{sp3}-Cl	199.1	47.7	300.0	201.8
C_{sp3}-Br	126.6	5.5	300.0	147.5

* Conjugated bond.

This shows that the values that are produced are far closer to those used by other force fields than those produced by the Chapter 7 final optimisation. Although this result is encouraging it is important to remember that in the Chapter 7 study the large shift in the parameters used to calculate bond strength only occurred during the final optimisation which involved far more optimisations than the final optimisation in this case. Though that study also showed that the change in total residual factor was small over the iterations where this optimisation took place, indicating once more the small effect that these parameters have on the residual factors.

8.5.3.3 Strain Free Valence Angles

As with these parameters it is difficult, especially from the weights, to determine that the parameters are reasonable. For the angles it is important to remember that the final optimisation for this study actually took place before the final optimisation for the structure set in Chapter 7. At this time during the programs development the total angles parameters were not fixed and it can be seen that several of the total angles have moved to what would normally be considered as unreasonable values. Main amongst these are the value for S_{Vi} which has gone out to nearly 880° and the C_{sp2} value which has moved out from 360° to over 390°. The likely reason why this happened is given in section 7.2.2. It was because of this problem, as it occurred, here that the values for these parameters were fixed in the

final optimisation in Chapter 7. Section 7.2.2 also stated that this problem would mainly be driven by the structural measures at the expense of the energy measures and it is quite likely that these large angular values are responsible for the fact that the structural measures improved so much at the expense of the energy measures. Unfortunately, due to the disk crash at the time this effect was coming to light, it was not possible to reoptimise the force field with these parameters fixed to more reasonable values. If it had been possible the experiment using different total angles for C_{sp2} in minimising Acetone that took place in section 7.2.2 indicates that effect on the final optimised structural properties would not be as dramatic as might be expected from the large deviations from the expected total angle values.

8.5.3.4 Angle Bend Force Constant Calculation Values

The values for these parameters are much closer to the values from CHEMMIN. Because the number of force field loops for this study was much less than that in Chapter 7 and that the drift of the hydrogen scaling factors was obviously a very slow shift across the final optimisation in that case it is difficult to say if these better parameters are due to the larger structure set or that the parameters would, given enough time, go out to the same sort of range. It is important to note that even with the current values of these parameters that the distribution of the valence angle residual factors for this study are very similar to that for the chapter 7 study so these values could be considered to be as, or possibly even more, reliable as those produced in that study.

8.5.3.5 Non-Bonded Parameters

Table 8.5 below shows the van der Waals Radii from both this study and from Cosmic. Along with these it lists the number of structures in the structure set which contained that particular atom type. It is possible to see that there is a distinct connection between the number of structures in which an atom type is found and the closeness of its van der Waals radii to standard values. This is shown very clearly by the fact that for the 3 groups in table 8.4 the average number of structures are: up to 0.1Å 82.7, 0.1Å to 0.3Å 29.6 and over 3Å 16.2. This shows once more that the approach used to optimise parameters in OPTPARAM is viable and that the main problem with using it to produce a standard force field is obtaining enough data of the correct type and the obvious slowing of the optimisation that will occur as the number of structures used increases.

Table 8.5 Non-Bonded Parameters and Occurrence of Atom Types

Within 0.1 Angstroms			
Atom Type	This Study	Cosmic	No.in Set
H	1.19	1.2	210
C_{sp2}	1.66	1.7	104
C_{sp3}	1.70	1.7	124
N_{amide}	1.51	1.55	45
N_{nitro}	1.56	1.55	14
O_{sp3}	1.44	1.52	59
F	1.42	1.47	23
Within 0.3 Angstroms			
Atom Type	This Study	Cosmic	No.in Set
C_{arom}	1.96	1.7	32
N_{arom}	1.75	1.55	4
O_{sp2}	1.34	1.52	102
$S_{ii=}$	2.00	1.80	16
Cl	1.89	1.75	32
Br	2.15	1.85	17
I	2.21	1.98	4
Outside 0.3 Angstroms			
Atom Type	This Study	Cosmic	No.in Set
C_{sp}	1.02	1.7	18
N_{sp}	2.29	1.55	17
N_{sp2}	1.17	1.55	20
N_{sp3}	1.18	1.55	15
N_{cation}	2.18	1.55	23
S-ii-	0.79	1.80	28
S_{iv}	1.23	1.80	1
S_{vi}	2.51	1.80	8

For the *f* factors the differences are even smaller. As COSMIC has only one *f* factor for each element rather than each atom type table 8.6 gives the COSMIC value as the second value for the first atom type associated with each element.

Table 8.6 Comparison of the Non-Bonded *f* Factors

Hydrogen		
Atom Type	This Study	Cosmic
H	0.0412	0.042
Carbon		
Atom Type	This Study	Cosmic
C_{sp}	0.1243	0.107
C_{sp2}	0.1333	
C_{arom}	0.1080	
C_{sp3}	0.1035	
Nitrogen		
Atom Type	This Study	Cosmic
N_{sp}	0.1180	0.095
N_{sp2}	0.0910	
N_{arom}	0.1071	
N_{amide}	0.0966	
N_{nitro}	0.0972	
N_{sp3}	0.0983	
N_{cation}	0.1055	
Oxygen		
Atom Type	This Study	Cosmic
O_{sp2}	0.1227	0.116
O_{sp3}	0.1131	
O_{anion}	0.1234	

Fluorine		
Atom Type	This Study	Cosmic
F	0.1051	0.109
Sulphur		
Atom Type	This Study	Cosmic
S-ii-	0.2659	0.314
$S_{ii=}$	0.1835	
S_{iv}	0.2789	
S_{vi}	0.3170	
Chlorine		
Atom Type	This Study	Cosmic
Cl	0.2927	0.314
Bromine		
Atom Type	This Study	Cosmic
Br	0.3140	0.434
Iodine		
Atom Type	This Study	Cosmic
I	0.6045	0.623

All Parameters have units of:
Kcal Mol^{-1}

The closeness of most of these values to those from Cosmic is very encouraging as it the energy parameters such as these which proved far harder to optimise than the more physically related parameters such as bond lengths.

8.5.3.6 Torsional Barriers

Table 8.7 below shows a selection of the torsional parameters from this study and compares these with the same parameters from both the study in chapter 7, the Cosmic force field and the White-Bovill force field WBFF[72]. Note that the WBFF has several torsion barriers for each set of central atom types depending on the other atoms types involved in the torsion angle.

The table shows that the values for this study are somewhat higher than those from the other studies. This is possibly a result of the technique used in this study of optimising all the parameter together rather than doing the parameters in sections of different atom types whilst keeping previous parameters fixed. This point will be covered in greater detail in the discussion.

Table 8.7 Optimised Torsion Barrier Parameters

Torsional Barriers	This Study	Chapter 7	Cosmic	WBFF
All Kcal Mol^{-1}				
Csp2=Csp2	8.125	5.737006	6.0	6.2500
Cps2-Cps3	0.182	0.086213	0.06	0.1367
				0.0629
Csp3-Csp3	0.133	0.078062	0.10	0.1100
				0.0200
				0.0629
Csp2-Csp2*	0.538	0.812628	-------	--------

*Conjugated

8.5.3.7 Out of Plane Bending Force Constants

Table 8.8 shows a comparison of the out of plane bending parameters. This shows that the values are very close to those from CHEMMIN from which they initially started.

Table 8.8 Optimised Out of Plane Bending Parameters

Out of Plane Bending	This study	Chapter 7	CHEMMIN
C_{sp2}	0.001208	0.002475	0.0012
C_{arom}	0.001160		0.0012
N_{amide}	0.000189		0.0002
N_{nitro}	0.000213		

All Kcal $Mole^{-1}$ Deg^{-2}

8.6 Discussion

Although the optimisations detailed in this study were not as extensive as those in the chapter 7 investigation it can be seen from figures 8.7 to 8.11 that the distribution of residual factors between the final runs in both cases are almost identical. When the shape of the typical optimisation curves, such as those given in chapter 7, are considered it is likely that the structural and thermochemical properties values given above will be close to those that would be produced had the optimisation proceeded further. Whether the optimised parameters are close to the values that they would have with further optimisation is rather more uncertain but as the final results using those, unknown, parameters would be close to those using the above parameters it is valid to use the above parameters as a valid force field.

Examining the various sections within the force field it can be seen that compared to the chapter 7 study more of the values could be considered as chemically reasonable. The worst case are the total angles used in the calculation of the strain free valence angles. It was as a direct result of the bad values that the total angles were fixed in the final optimisation in chapter 7. The fact that the distribution of the residual factors in that study are slightly better than those in this one indicates that had this study been done with fixed total angles the results would only have been slightly worse.

Of the other parameter sections that are further from what might be expected the most obvious are the van der Waals radii detailed in section 8.5.3.5. As mentioned there the deviation from the expected range for these values is closely linked to the number of structures in which each atom type is present. The direct implication of this is that the method of producing parameters in OPTPARAM is viable and that any artefacts within the produced parameters are due to deficiencies in the structure set, which can be correctable, rather than in the technique that is used.

Another point to be noted about both this optimisation and that in chapter 7 is that unlike other force field optimisations all the parameters are optimised at the same time. This compares with other methods where only those parameters for new atom types are optimised while those for the remaining atom types remain fixed. The advantage of the method used here can be seen by considering the graph in figure 7.8. If this bond length had been optimised for a force field without electronegative elements such as halogens then the use of that parameter within a force field that included halogens would be less effective than had the parameter been optimised by the method used in this study. The same is likely to be true for torsional barriers where with the 'central two' parameter values for an alkene force

field are likely to be different from the same parameters when optimised with a more general set of structures.

In conclusion the investigation in this study has shown than the use of a highly reduced force field can produce reasonable results and that given sufficient structural and thermochemical data it is possible to automate the process of parameter optimisation with only the minimum amount of intervention.

8.7 References

1 F.H.Allen, O.Kennard and R.Taylor, *Acc. Chem. Res.*, 1983, **16**, 146.

2 M.J.S.Dewer and W.Thiel, *J. Am. Chem. Soc.*, 1977, **99**, 4899.

3 CHARGE2, R.J.Abraham, G.H.Grant, B Hudson and P.E.Smith, University of Liverpool, Liverpool, UK.

4 Thermochemical Data of Organic Compounds, J.B.Pedley, R.D.Naylor and S.P.Kirby, Chapman and Hall Ltd., 1986, ISBN 0-412-27100-1.

5 J.R.Furig, S.E.Godbey and J.F.Sullivan, *J. Chem. Phys.*, 1984, **80**, 5983.

6 R.J.Abraham and G.Gatti, *J. Chem. Soc.* B, 1969, 961.

7 D.Freisen and K.Hedburg, *J. Am. Chem. Soc.*, 1980, **102**, 3982.

8 L.Fernholt and K.Kveseth. *Acta. Chem. Scand.*, 1978, **A32**, 63.

9 K.Kveseth, *Acta. Chem. Scand.*, 1975, **A29**, 307.

10 D.Freisen and K.Hedburg, *J. Am. Chem. Soc.*, 1980, **102**, 3982.

11 K.Hagen, K.Hedburg, J.Neisess and G.Gundersen, *J. Am. Chem. Soc.*, 1985, **107**, 341.

12 O.Trongmo, Q.Shen, K.Hagen and R.Seip, *J. Mol. Struc.*, 1981, **71**, 185.

13 O.I.Sovik, H.Schei, R.Stolevik, K.Hagen and Q.Shen. *J. Mol. Struc.*, 1984, **116**, 239.

14 Q.Shen, *J. Mol. Struc.*, 1984, **112**, 169.

15 T.Nordtomme and K.Hagen, *J. Mol. Struc.*, 1985, **128**, 127.

16 K.Hagen, *J. Mol. Struc.*, 1985, **128**, 139.

17 A.L.Verma, U.F.Murphy and H.J.J.Bernstein, *J. Chem. Phys.*, 1974, **60**, 1540.

18 E.N.Marvell and R.S.Knutson, *J. Org. Chem.*, 1970, **35**, 338.

19 O.Steinnes, Q.Shen and K.Hagen, *J. Mol. Struc.*, 1980, **64**, 217.

20 G.Gunderson, H.G.Thomassen, J.E.Boggs and M.L.Collins. *J. Mol. Struc.*, 1989, **213**, 1.

21 F.Kapteijn, A.J. van der Steen and J.C.Mol, *J. Chem. Thermodyn.*, 1983, **15**, 137.

22 M.Davis and D.Hassel, *Acta Chem. Scand.*, 1963, **17**, 1181.

23 M.Traetteberg, P.Bakken, R.Seip, S.J.Cyvin, B.N.Cyvin and H.Hopf,*J. Mol. Struc.*, 1985, **51**, 77.

24 M.Traetteberg, P.Bakken, R.Seip, S.J.Cyvin, B.N.Cyvin and H.Hopf,*J. Mol. Struc.*, 1985, **55**, 199.

25 I.F.Shishkov, N.I.Sadove, L.V.Vilkov and Yu.A.Pankrushev, *J. Struc. Chem.*, 1983, **24**, 25.

26 W.Penbrinker, P.A.G.Huisman and F.C.Mijhoff, *J. Mol. Struc.*, 1979, **54**, 293.

27 S.H.Schei, *J. Mol. Struc.*, 1983, **98**, 141.

28 G.Schultz, I.Hargittai, M.Kolonits and J.Garbarczyk, *J. Mol. Struc.*, 1987, **160**, 267.

29 V.P.Novikov, M.V.Popik, L.V.Vilkov, G.I.Migachev and K.M.Dyumaev, *J. Mol. Struc.*, 1985, **53**, 211.

30 A.L.Andreassen and S.H.Bauer, *J. Mol. Struc.*, 1972, **12**, 381.

31 O.E.Kjeldseth Moe and T.G.Strand, *J. Mol. Struc.*, 1985, **128**, 13.

32 A.Almenningen, O.Bastiansen, L.Fernholt, B.N.Cyvin,S.J.Cyvin and S.Samdel, *J. Mol. Struc.*, 1985, **128**, 59.

33 D.W.H.Rankin, H.E.Robertson and H.H.Karsh, *J. Mol. Struc.*, 1985, **77**, 121.

34 D.W.H.Rankin, M.R.Todd and M.Fild, *J. Chem. Soc. Dalton.*, 1985, **82**, 2079.

35 H.Oberhammer, G.D.Knerr and J.M.Shreeve, *J. Mol. Struc.*, 1985, **82**, 143.

36 V.Typke, M.Dakkouri and F.Schlumberger, *J. Mol. Struc.*, 1980, **62**, 111.

37 H.Oberhammer, W.Gomber and H.Willner, *J. Mol. Struc.*, 1985, **70**, 273.

38 A.Tsuboyama, S.Konaka and M.Kimura, *J. Mol. Struc.*, 1985, **127**, 77.

39 L.Acha, E.R.Crombie and D.W.H.Rankin, *J. Mol. Struc.*, 1985, **73**, 111.

40 K.Kimura and M.Kubo, *J. Chem. Phys.*, 1959, **30**, 151.

41 E.E.Astrop, *Acta. Chem. Scand.*, 1973, **27**, 3271.

42 B.Haas and H.Oberhammer, *J. Am. Chem. Soc.*, 1984, **106**, 6146.

43 G.Portalone, A.Dominicano, G.Schultz and I.Hargittai, *J. Mol. Struc.*, 1987, **160**, 97.

44 B.Rozsondai, J.H.Moore, D.C.Gregory and I.Hargittai, *J. Mol. Struc.*, 1985, **51**, 69.

45 J.Brunnell, K.Klonits, C.Bliefert, K.Seppelt and I.Hargittai, *J. Mol. Struc.*, 1985, **78**, 307.

46 C.Kato, S.Konaka, T.Iijima and M.Kimura, *Bull. Chem. Soc. Japan*, 1969, **42**, 2148

47 H.Burger, G.Pawelke and H.Oberhammer, *J. Mol. Struc.*, 1985, **128**, 283.
48 B.Beagley, S.P.Narola and V.Ulbrecht, *J. Mol. Struc.*, 1985, **56**, 207.
49 O.G.Batyokhnova, N.I.Sadova, L.V.Vilkov and Yu.A.Pankrushev,, *J. Mol. Struc.*, 1985, **97**, 153.
50 A.Almenningen, J.Brunvell, M.V.Popik, S.V.Sokolkov, L.V.Vilkov and S.Samdal, *J. Mol. Struc.*, 1985, **127**, 85.
51 G.Portalone, G.Schultz, A.Domenicano and I.Harittai, *J. Mol. Struc.*, 1985, **118**, 53.
52 D.W.H.Rankin, M.R.Todd, F.G.Riddell and S.Turner, *J. Mol. Struc.*, 1981, **71**, 171.
53 Z.Nahlovska, B.Nahlovsky and T.G.Strand, *Acta Chem. Scand.*, 1970, **24**, 2617.
54 S.N.Ketkar and M.Fink, *J. Mol. Struc.*, 1985, **77**, 139.
55 A.Almenningen, J.Brunvell, M.V.Popik, S.V.Sokolkov, L.V.Vilkov and S.Samdal,*J. Mol. Struc.*, 1985, **118**, 37.
56 J.Brunvell, R.K.Bohn and I.Hargittai, *J. Mol. Struc.*, 1985, **189**, 81.
57 E.J.Jacob, D.D.Danielson and S.Samdel, *J. Mol. Struc.*, 1985, **62**, 143.
58 M.Hargittai and I.Hargittai, *J. Mol. Struc.*, 1985, **73**, 253.
59 V.A.Navmov, O.A.Litvinov, H.J.Geise and J.Dillen, *J. Mol. Struc.*, 1983, **99**, 303.
60 L.Fernholt, S.Samdal and R.Seip, *J. Mol. Struc.*, 1981, **72**, 217.
61 D.Christen, *J. Mol. Struc.*, 1985, **66**, 211.
62 D.Christen, H.Oberhammer, W.Zeil, A.Haas and A.Darmadi, *J. Mol. Struc.*, 1985, **66**, 203.
63 J.Brunvell, I.Hargittai, T.Szekely and G.C.Pappardo, *J. Mol. Struc.*, 1985, **66**, 173.
64 O.G.Batyukhnova, N.I.Sadova, L.V.Vilcov, V.P.Ivshin and Yu.A.Pankrushev, *J. Struc. Chem.*, 1984, **25**, 878.
65 H.Schei, Q.Shen and R.L.Hilderbrandt, *J. Mol. Struc.*, 1980, **65**, 297.
66 M.Nakata, T.Fukuyama, and K.Kuchitsu, *J. Mol. Struc.*, 1985, **81**, 121.
67 J.Bronvell, O.Exner, I.Hargittai, M.Hargittai and P.Scharfenburg, *J. Mol. Struc.*, 1985, **117**, 311.
68 A.Almenningen, P.Kolsaker, H.M.Seipe and T.Willadsen, *Acta. Chem. Scand.*, 1969, **23**, 3398.
69 C.J.Marsden, *Inorg. Chem.*, 1984, **23**, 1703.
70 J.G.Vinter, A.Davis and M.R.Saunders, *J.Comput.-Aided Mol. Design*, 1987, **1**, 31.

71 D.N.J.White, J.N.Ruddock and P.R.Edgington, *Molecular Simulation*, 1989, **3**, 71.

72 D.N.J.White and M.J.Bovill, *J. Chem. Soc. Perkins*, 1977, **II**, 1610.

APPENDIX B Typical COMPARE file

```
COMPARE FILE  DATE :- 040490

CIS BUT 2 ENE                                  *
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 DIST   0  1           1.504     1.506 +/- 0.002   0.000
 DIST   1  2           1.342     1.346 +/- 0.003   0.001  <>
 ANGL   0  1  2       127.3     125.4 +/-   0.4     1.5  <###>

TRANS BUT 2 ENE                                *
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 DIST   0  1           1.503     1.508 +/- 0.002   0.003  <#>
 DIST   1  2           1.341     1.347 +/- 0.003   0.003  <#>
 ANGL   0  1  2       124.6     123.8 +/-   0.4     0.4  <>

ADAMANTANE
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 DIST   0  1           1.539     1.534 +/- 0.004   0.001  <>
 ANGL   0  1  4       110.2     110.0 +/-   0.5     0.0
 ANGL   1  0  2       109.1     109.2 +/-   0.5     0.0

BICYCLO[2.2.2]OCT-2-ENE
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 TORS   3  0  7  4     122.5     121.2 +/-   2.0     0.0
 TORS   0  1  4  7       0.0       0.0 +/-   2.0     0.0

BICYLO[2.2]OCTA-2,5-DIENE
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 TORS   1  0  7  5     121.2     123.4 +/-   2.0     0.2  <>
 TORS   0  2  5  7       0.0       0.0 +/-   2.0     0.0

CIS-BUT 1 ENE
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 ANGL   0  1  2       127.1     126.7 +/-   0.5     0.0
 ANGL   1  2  3       116.3     114.8 +/-   0.5     1.0  <#>
 TORS   0  1  2  3       0.0       0.0 +/-   2.0     0.0

CYCLODECA-1,6,-DIENE
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 DIST   2 11           1.109     1.112 +/- 0.004   0.000
 DIST   4  5           1.342     1.326 +/- 0.004   0.012  <##>
 DIST   1  2           1.537     1.534 +/- 0.006   0.000
 DIST   3  4           1.507     1.506 +/- 0.006   0.000
 ANGL   3  4  5       127.7     128.2 +/-   0.3     0.2  <>
 ANGL   2  3  4       112.3     112.8 +/-   0.3     0.2  <>
 ANGL   1  2  3       114.0     114.1 +/-   0.5     0.0
 ANGL  23  0 24       117.9     116.6 +/-   1.0     0.3  <>
 ANGL  11  2 12       106.6     105.6 +/-   1.0     0.0  <>

CYCLOHEXA-1,4-DIENE
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 DIST   5 12           1.108     1.114 +/- 0.003   0.003  <#>
 DIST   0  6           1.097     1.103 +/- 0.003   0.003  <>
 DIST   0  1           1.339     1.334 +/- 0.002   0.003  <#>
 DIST   0  5           1.504     1.496 +/- 0.002   0.006  <##>
 ANGL   1  2  3       113.4     113.3 +/-   0.3     0.0
 ANGL   2  3  4       123.3     123.4 +/-   0.2     0.0
 ANGL   1  0  6       120.3     123.4 +/-   2.7     0.4  <>
 ANGL   0  5 12       109.4     110.0 +/-   0.4     0.2  <>
 ANGL  12  5 13       105.5     103.0 +/-   2.0     0.5  <>
 ANGL   5  0  6       116.4     113.2 +/-   2.7     0.5  <>

CYCLOHEXENE
 TYPE   ATOM NO'S     CALC     EXPT    ERROR       DIF
-------------------------------------------------------
 DIST   0  1           1.340     1.335 +/- 0.003   0.002  <>
 DIST   1  2           1.505     1.504 +/- 0.006   0.000
 DIST   2  3           1.536     1.515 +/- 0.020   0.001  <>
 DIST   0  6           1.097     1.093 +/- 0.015   0.000
 ANGL   1  0  5       123.5     123.5 +/-   0.5     0.0
 ANGL   0  5  4       111.9     112.1 +/-   0.5     0.0
 ANGL   2  3  4       110.7     110.0 +/-   0.5     0.2  <>
 TORS   0  1  2  3      15.6      15.2 +/-   2.0     0.0
 TORS   1  2  3  4     -45.1     -44.9 +/-   2.0     0.0
 TORS   2  3  4  5      61.2      60.2 +/-   2.0     0.0
```

CYCLOPENTENE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
TORS	0 4 2 3	162.7	156.7	+/- 2.0	4.0	<#>

2,3-DIMETHYLBUT-2-ENE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
DIST	0 1	1.350	1.353	+/- 0.004	0.000	
DIST	0 2	1.512	1.511	+/- 0.002	0.000	
ANGL	0 1 4	123.0	123.9	+/- 0.5	0.4	<>

ETHYLENE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
DIST	0 1	1.337	1.335	+/- 0.003	0.000	
DIST	0 3	1.097	1.090	+/- 0.003	0.004	<#>
ANGL	0 1 2	121.3	121.7	+/- 0.4	0.0	<>

ISOBUTENE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF
ANGL	0 1 2	122.2	122.1	+/- 0.3	0.0

MANXANE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
DIST	15 20	2.241	2.200	+/- 0.020	0.021	<#>

2-METHYL-SKEW-BUT-1-ENE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
TORS	2 1 3 4	84.7	72.7	+/- 2.5	9.5	<###>

NORBORNADIENE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
DIST	0 1	1.506	1.535	+/- 0.007	0.022	<###>
DIST	1 4	1.338	1.343	+/- 0.003	0.002	<>
DIST	0 3	1.535	1.544	+/- 0.004	0.005	<#>
ANGL	0 3 6	94.9	94.1	+/- 3.0	0.0	
TORS	1 0 6 5	115.7	115.6	+/- 2.0	0.0	

PIN-2-ENE *

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
DIST	0 1	1.344	1.340	+/- 0.010	0.000	
ANGL	0 4 5	108.7	112.0	+/- 3.0	0.3	<>
ANGL	0 1 3	115.8	118.0	+/- 3.0	0.0	
ANGL	1 0 4	118.7	118.0	+/- 3.0	0.0	
ANGL	0 1 2	125.4	126.0	+/- 3.0	0.0	
TORS	6 5 3 7	140.0	146.0	+/- 8.0	0.0	

PROPENE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
DIST	0 1	1.339	1.341	+/- 0.001	0.001	<#>
DIST	1 2	1.503	1.504	+/- 0.001	0.000	<>
ANGL	0 1 2	124.7	124.8	+/- 0.1	0.0	<>

SKEW-BUT-1-ENE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
ANGL	0 1 2	124.3	125.4	+/- 0.5	0.6	<#>
ANGL	1 2 3	111.3	112.1	+/- 0.5	0.3	<>

TRANS,TRANS,TRANS,CYCLODODECA-1,5,9-TRIE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
DIST	1 2	1.535	1.540	+/- 0.010	0.000	
DIST	2 3	1.505	1.490	+/- 0.010	0.005	<>
DIST	3 4	1.340	1.320	+/- 0.010	0.010	<>
ANGL	1 2 3	111.3	111.1	+/- 0.5	0.0	
ANGL	2 3 4	124.1	124.1	+/- 0.5	0.0	
TORS	0 1 2 3	61.8	63.4	+/- 2.0	0.0	
TORS	1 2 3 4	-115.8	-116.5	+/- 2.0	0.0	
TORS	2 3 4 5	179.0	178.0	+/- 2.0	0.0	

CYCLOHEXANE

TYPE	ATOM NO'S	CALC	EXPT	ERROR	DIF	
DIST	0 1	1.535	1.528	+/- 0.004	0.003	<>
DIST	0 6	1.108	1.100	+/- 0.004	0.004	<#>
ANGL	0 1 2	111.1	111.3	+/- 0.5	0.0	
ANGL	6 0 7	108.1	110.0	+/- 2.5	0.0	
TORS	0 1 2 3	55.9	55.2	+/- 0.5	0.2	<>

ENERGY DIFFERENCES

BETWEEN GAUCHE-BUTANE *
AND ANTI-BUTANE *

CALC 0.663 EXPT 0.824 ERROR 0.143 DIF 0.018

BETWEEN CIS BUT 2 ENE *
AND TRANS BUT 2 ENE *

CALC 1.509 EXPT 1.130 ERROR 0.380 DIF 0.000

BETWEEN BOAT-CHAIR BICYCLO[3.3.1] NONANE *
AND twin-chair bicyclo[3.3.1] nonane *

CALC 2.841 EXPT 3.050 ERROR 0.150 DIF 0.059

CIS BUT 2 ENE *
CALC : -1.2572 EXPT : -1.6970 ERROR : 0.2400 DIF : 0.1998
TRANS BUT 2 ENE *
CALC : -2.7660 EXPT : -2.7250 ERROR : 0.2400 DIF : 0.0000
ADAMANTANE
CALC : -33.0511 EXPT : -32.1700 ERROR : 0.5500 DIF : 0.3311
ch-ch-bicyclo(3.3.2.)decane
CALC : -25.5853 EXPT : -25.3580 ERROR : 1.7690 DIF : 0.0000
cis-hex-2-ene
CALC : -11.9401 EXPT : -12.5100 ERROR : 0.3300 DIF : 0.2399
CYCLOHEXA-1,4-DIENE
CALC : 26.5708 EXPT : 26.3000 ERROR : 0.1000 DIF : 0.1708
CYCLOHEXENE
CALC : -0.4620 EXPT : -1.1950 ERROR : 0.1670 DIF : 0.5660
cycloheptene
CALC : -2.8101 EXPT : -2.1990 ERROR : 0.2630 DIF : 0.3481
CYCLOPENTENE
CALC : 9.2069 EXPT : 8.1020 ERROR : 0.3350 DIF : 0.7699
2,3-dimythyl-but-1-ene
CALC : -15.2867 EXPT : -14.9620 ERROR : 0.3100 DIF : 0.0147
2,3-DIMETHYLBUT-2-ENE
CALC : -16.2485 EXPT : -16.3000 ERROR : 0.2630 DIF : 0.0000
4,4-dimethyl-pent-1-ene
CALC : -18.7033 EXPT : -19.5030 ERROR : 0.4780 DIF : 0.3217
ETHYLENE
CALC : 12.7909 EXPT : 12.5480 ERROR : 0.0960 DIF : 0.1469
hexa-1,5-diene
CALC : 20.4412 EXPT : 20.1000 ERROR : 0.2150 DIF : 0.1262
ISOBUTENE
CALC : -4.5253 EXPT : -4.0390 ERROR : 0.2150 DIF : 0.2713
1-methyl-adamantane
CALC : -39.9916 EXPT : -41.0130 ERROR : 0.6690 DIF : 0.3524
1-methyl-cyclohexene
CALC : -9.5468 EXPT : -10.3490 ERROR : 0.1910 DIF : 0.6112
1-methyl-cyclopentene
CALC : -0.5435 EXPT : -0.9080 ERROR : 0.1670 DIF : 0.1975
1-methyl-diamantane
CALC : -40.7923 EXPT : -39.8420 ERROR : 0.7890 DIF : 0.1613
2-methyl-but-2-ene
CALC : -10.1261 EXPT : -9.9900 ERROR : 0.2630 DIF : 0.0000
2-methyl-pent-2-ene
CALC : -15.0641 EXPT : -15.9900 ERROR : 0.3590 DIF : 0.5669
3-methyl-but-1-ene
CALC : -6.7103 EXPT : -6.5970 ERROR : 0.1910 DIF : 0.0000
4-methyl-cis-pent-2-ene
CALC : -14.4816 EXPT : -13.7430 ERROR : 0.2870 DIF : 0.4516
4-methyl-trans-pent-2-ene
CALC : -14.5164 EXPT : -14.6990 ERROR : 0.3350 DIF : 0.0000
2-methyl-but-1-ene
CALC : -8.8535 EXPT : -8.4370 ERROR : 0.2390 DIF : 0.1775
NORBORNADIENE
CALC : 59.3419 EXPT : 58.7720 ERROR : 0.6930 DIF : 0.0000
norbornane
CALC : -11.8823 EXPT : -13.1210 ERROR : 1.1230 DIF : 0.1157
pent-1-ene
CALC : -5.3102 EXPT : -5.0910 ERROR : 0.2390 DIF : 0.0000

```
PROPENE
CALC :    4.7979  EXPT :    4.7800  ERROR :    0.1910  DIF :    0.0000
penta-1,4,diene
CALC :   25.4421  EXPT :   25.2390  ERROR :    0.2870  DIF :    0.0000
peritrihydroquinacene
CALC :  -23.4695  EXPT :  -24.4500  ERROR :    0.8370  DIF :    0.1435
5-skew-cis-pent-2-ene
CALC :   -6.6492  EXPT :   -6.8600  ERROR :    0.2000  DIF :    0.0108
5-skew-trans-pent-1-ene
CALC :   -7.8066  EXPT :   -7.7600  ERROR :    0.2000  DIF :    0.0000
TRANS,TRANS,TRANS,CYCLODODECA-1,5,9-TRIE
CALC :   22.5160  EXPT :   24.2110  ERROR :    1.0280  DIF :    0.6670
CYCLOHEXANE                                      *
CALC :  -30.3150  EXPT :  -29.5000  ERROR :    0.2000  DIF :    0.6150
trans-hex-2-ene
CALC :  -13.0455  EXPT :  -12.8800  ERROR :    0.3900  DIF :    0.0000
1,3,5,7-tetramethyl-adamantane
CALC :  -67.7198  EXPT :  -67.7340  ERROR :    1.6490  DIF :    0.0000
2,4,4-trimethyl-pent-1-ene
CALC :  -27.2660  EXPT :  -26.4340  ERROR :    0.3580  DIF :    0.4740

 STRUCTURE SUM     38.952


AVERAGE BOND LENGTH DIFFERENCE          :-   0.0079    AVERAGE ERROR  :-   0.0056

AVERAGE BOND ANGLE DIFFERENCE           :-   0.9211    AVERAGE ERROR  :-   1.0571

AVERAGE TORSION ANGLE DIFFERENCE        :-   2.0722    AVERAGE ERROR  :-   2.3125

AVERAGE CONFORMER DIFFERENCE            :-   0.2494    AVERAGE ERROR  :-   0.2243

AVERAGE HEAT OF FORMATION DIFFERENCE    :-   0.5207    AVERAGE ERROR  :-   0.4331

AVERAGE OF SUM 1st    :-   0.0178

STRUCTURE DIFFERENCE :-   0.5923
CONFORMER ENERGY     :-   0.0037
HEAT OF FORMATION    :-   0.0505

TOTAL                :-   0.6466

AVERAGE OF SUM 1st    :-   0.0178

NUMBER OF PARAMETERS USED :-67

 ITERATIONS 20, DATE 040490, SBIAS   1.0, EBIAS   1.0, HBIAS   1.0

 FILES USED FOR THIS OPTIMISATION

 BOND LENGTH , FULL , NUMBER 89
 BOND STRENGTH , FULL , NUMBER 22
 NON BONDED , COSMIC TYPE , NUMBER 61
 ANGLE VALUES , FULL WITH SUB FACTOR , NUMBER 18
 ANGLE BEND , FULL , NUMBER 48
 TORSION , FULL (as per WBFF), NUMBER 27
 OUT OF PLANE BENDING , IMPLEMENTED , NUMBER 89

 STRUCTURE          :      0.25028
 ENERGY             :      0.00375
 HEAT OF FORMATION  :      0.05053

 TOTAL              :      0.30455
```

Appendix C Typical Parameter Information File

```
ITERATIONS 100, DATE 201290, SBIAS   5.0, EBIAS  20.0, HBIAS   1.0

FILES USED FOR THIS OPTIMISATION

BOND LENGTH , FULL , NUMBER 1
BOND STRENGTH , CALCULATED , NUMBER 1
NON BONDED , COSMIC TYPE , NUMBER 1
ANGLE VALUES , CALCULATED , NUMBER 1
ANGLE BEND , CALCULATED , NUMBER 1
TORSION , CENTRAL TWO , NUMBER 1
OUT OF PLANE BENDING , IMPLEMENTED , NUMBER 1

STRUCTURE          :     10.39974
ENERGY             :      0.11543
HEAT OF FORMATION  :      0.44592

TOTAL              :     54.75324
```

Appendix D Typical STRUCTUR.LST file

```
C:\OPTPARAM\MOLDATA\TEST.DAT
C:\OPTPARAM\MOLDATA\GBUTANE.DAT
C:\OPTPARAM\MOLDATA\ABUTANE.DAT
C:\OPTPARAM\MOLDATA\CBUT2ENE.DAT
C:\OPTPARAM\MOLDATA\TBUT2ENE.DAT
C:\OPTPARAM\MOLDATA\CCBCYNON.DAT
C:\OPTPARAM\MOLDATA\BCBCYNON.DAT
C:\OPTPARAM\MOLDATA\PIN2ENE.DAT
C:\OPTPARAM\MOLDATA\BCOC2ENE.DAT
C:\OPTPARAM\MOLDATA\CCD16IEN.DAT
C:\OPTPARAM\MOLDATA\ADMNTANE.DAT
C:\OPTPARAM\MOLDATA\PROPENE.DAT
C:\OPTPARAM\MOLDATA\ETHYLENE.DAT
C:\OPTPARAM\MOLDATA\BCOCDIEN.DAT
C:\OPTPARAM\MOLDATA\DIMYHB2N.DAT
C:\OPTPARAM\MOLDATA\CYPNTENE.DAT
C:\OPTPARAM\MOLDATA\CYHEXENE.DAT
C:\OPTPARAM\MOLDATA\NBRNIENE.DAT
C:\OPTPARAM\MOLDATA\TCYDODTR.DAT
C:\OPTPARAM\MOLDATA\MANXANE.DAT
C:\OPTPARAM\MOLDATA\ISBUTENE.DAT
C:\OPTPARAM\MOLDATA\CBUT1ENE.DAT
C:\OPTPARAM\MOLDATA\CHX14IEN.DAT
C:\OPTPARAM\MOLDATA\SBUT1ENE.DAT
END
3
   2    3    0.824 0.143
   4    5    1.130 0.380
   7    6    3.050 0.150
```

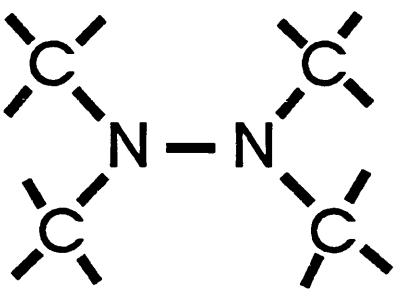
Appendix E Typical Structure File

```
ACETONE
    1.000     1.000     1.000    90.000    90.000    90.000
  10
C     1       -1.2296   -0.5186    0.0002    -0.07924    1
C     2        0.2973   -0.5330   -0.0008     0.18800    1
C     1        1.0567    0.7917   -0.0001    -0.07924    1
O    11        0.8999   -1.5845   -0.0021    -0.34151    1
H    15       -1.6872   -1.5107   -0.0004     0.05200    1
H    15       -1.6183    0.0004   -0.8778     0.05200    1
H    15       -1.6173   -0.0011    0.8795     0.05200    1
H    15        2.1441    0.6850   -0.0008     0.05200    1
H    15        0.8064    1.3879    0.8792     0.05200    1
H    15        0.8053    1.3894   -0.8780     0.05200    1
  10
   1   2   5   6   7   0   0
   2   1   3   4   0   0   0
   3   2   8   9  10   0   0
   4   2   0   0   0   0   0
   5   1   0   0   0   0   0
   6   1   0   0   0   0   0
   7   1   0   0   0   0   0
   8   3   0   0   0   0   0
   9   3   0   0   0   0   0
  10   3   0   0   0   0   0
 .1000E-02
   5
D
   1   2             1.517   .003
D
   2   4             1.210   .004
D
   1   5             1.091   .003
A
   6   1   5         108.5    .5
A
   1   2   3         116.0    .3
   1   2
   2   3
   1  15
   0   0
-51.90     .17
```

APPENDIX F Typical Detailed Data Sheet.

Reference Number :- 91
Molecule Name :- **TETRAMETHYL HYDRAZINE**
Full Reference :- V.A.NAVMOV, O.A.LITINOV, H.J.GEISE & J.DILLON J. Mol. Struc. 99 (1983) 303

①	2	3	4	5	6	⑦	8	9	10	⑪	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40					



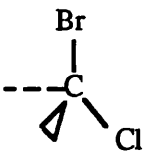
C-H	1.096(4)
C-N	1.463(1)
N-N	1.401(4)
N-C-H	108.6(7)
C-N-C	110.8(18)
N-N-C	113.5(6)
C-N-N-C	78.5(41)

Appendix G Typical Initial Survey Form

Ref Number :- **53**

Name:- **BROMOCHLORO METHANE** Reference :- **J.Mol.. Struc., 125, 143**

Formula :-



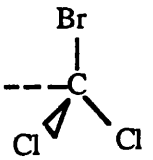
Number of conformers :- **1** Deformed :- **N**

Energy Difference Known [y/n] :- **N**

Ref Number :- **54**

Name:- **BROMODICHLORO METHANE** Reference :- **J.Mol.. Struc., 125, 143**

Formula :-



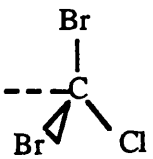
Number of conformers :- **1** Deformed :- **N**

Energy Difference Known [y/n] :- **N**

Ref Number :- **55**

Name:- **DIBROMOCHLORO METHANE** Reference :- **J.Mol.. Struc., 125, 143**

Formula :-



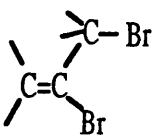
Number of conformers :- **1** Deformed :- **N**

Energy Difference Known [y/n] :- **N**

Ref Number :- **56**

Name:- **2,3 DIBROMO-1-PROPENE** Reference :- **J.Mol.. Struc., 118,1**

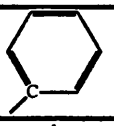
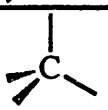
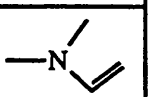
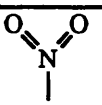
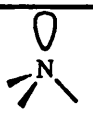
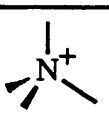
Formula :-

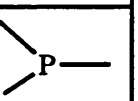
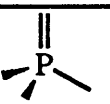
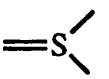


Number of conformers :- **2** Deformed :- **N**

Energy Difference Known [y/n] :- **Y**

Appendix H Typical Atom Types Required for a Full Force Field

H	–H
C_{sp}	–C≡
C_{sp2}	=C<
C_{arom}	
C_{sp3}	
N_{sp}	≡N
N_{sp2}	
N_{arom}	
N_{amide}	
N_{nitro}	
N_{sp3}	
N_{cation}	
O_{sp2}	=O
O_{sp3}	–O–
O_{anion}	$-O^-$

F	–F
Na	Na^+
Mg	Mg^{2+}
P_{iii}	
P_v	
S_{ii} (a)	–S–
S_{ii} (b)	=S
S_{iv}	
S_{vi}	
Cl	–Cl
K	K^+
Ca	Ca^{2+}
Fe_{ii}	Fe^{2+}
Fe_{iii}	Fe^{3+}
Co_{ii}	Co^{2+}

Co_{iii}	Co^{3+}
Cu_i	Cu^+
Cu_{ii}	Cu^{2+}
Zn	Zn^{2+}
Br	–Br
I	–I
Ba_{ii}	Ba^{2+}